

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Application No.: 10/666,642 )  
In re application of: JIANG, Cai-Zhong, *et al.* )  
Filed: September 18, 2003 )  
Art Unit: 1638 )  
Examiner: BAUM, Stuart F. )  
Docket No. MBI-0054 )  
Customer No. 47334 )  
\_\_\_\_\_ )

Mail Stop Amendment  
Commissioner for Patents  
P.O. Box 1450  
Alexandria, VA 22313-1450

**DECLARATION UNDER 37 CFR 1.132 OF PETER REPETTI**

I, Peter Repetti, declare:

1. I received my Bachelor of Science degree in Plant Science from The Pennsylvania State University and my doctoral degree in Plant Biology from The University of California, Berkeley. I joined Mendel Biotechnology in June of 2002 and have served as Senior Scientist since January of 2004. In this declaration, I serve as expert witness in that my work has involved characterization and use of cloned plant genes for modifying a variety of traits in genetically transformed plants, specifically in the areas of developmental alterations and the regulation of environmental stress responses. I have contributed to and supervised research in the area of environmental stress tolerance of plants that ectopically express sequences of the present invention, and I am therefore familiar with the present invention.

2. I understand that this application relates to transgenic plants transformed with an expression vector encoding a polypeptide having a conserved domain that is 70% identical to the conserved domain of G1274, SEQ ID NO: 194, seed produced by these transgenic plants, methods for producing the transgenic plants, and methods for increasing plant biomass or the tolerance of a plant to an abiotic stress. The transgenic plants express plant transcription factor polypeptides first identified in *Arabidopsis thaliana*, a plant that is widely used as a model species.

3. For the purposes of this declaration, a plant “line” means the progeny (through seed or vegetative propagation) of a transformation event or a newly bred variety (specific genotype).

4. The present application provides methods for analysis and identification of sequences from diverse species that are closely-related. These methods include phylogenetic analysis, sequence alignments and percentage identity determination.

Exhibit A provides a number of sequences that are closely related to G1274 and fall within, or just outside of, the G1274 clade of transcription factor polypeptides. These sequences are derived from diverse species that include dicots *Arabidopsis thaliana* and *Glycine max* and monocots *Oryza sativa* and *Zea mays*. Some sequences within the clade are found in the large box in Exhibit D. These polypeptide sequences descend from a common ancestral sequence, represented by the node at arrow “a” in Exhibit D, and the G1274 clade is encompassed by the large box in Exhibit D. The G1274 clade polypeptides comprise conserved domains, indicated by the underlined residues in Exhibit A, that are at least 74% identical to the conserved domain of G1274, amino acid coordinates 111-164.

Sequences that lie just outside of the clade (for example, those derived from a somewhat more distant ancestral sequence at arrow “b”) also comprise conserved domains similar to that found in G1274, but these are less similar to G1274 than the sequences within the clade. For example, G179, G2517, G194, and G1758 have conserved domains, indicated by the respective underlined residues in Exhibit A, that are 55% to 61% identical to the conserved domain of G1274.

5. Exhibit B provides data obtained with plants overexpressing a number of G1274 clade member sequences. Even though some of these results were obtained with a limited number of lines, many of the lines of plants that overexpress these sequences that comprise conserved domains at least 55% identical to the G1274 conserved domain have been observed to confer improved traits similar to those found in G1274 overexpressors. These observations were based on positive results of morphological and physiological assays described in Exhibit C. These improved traits include increased tolerance to water deprivation, cold during germination, cold during growth, and/or low nitrogen conditions. Several of these sequences when overexpressed also conferred broader leaves, larger rosettes and/or larger seedlings relative to controls. Thus, sequences that are phylogenetically and closely related to G1274 confer increased biomass in overexpressing plants.

Based on the limited tests performed thus far, a minority of the sequences within the G1274 clade, which have conserved domains at least 74% identical to the conserved domain of G1274, including G3728, G3719, G3730, and G3723, has not yet conferred traits similar to those conferred by G1274 in morphological or physiological assays in a significant number of plants. However, as noted below, a small number of lines of plants overexpressing each of these sequences have thus far tested positive for improved traits similar to those conferred by G1274. For example:

some 35S::G3728 seedlings of one line were slightly more tolerant to desiccation than controls in a plate-based assay, and 4 of 60 lines tested produced larger seedlings than controls;

seedlings of one of four 35S::G3719 lines were larger and produced less anthocyanin than controls on low nitrogen media;

in a cold germination assay, one of four lines of G3730 overexpressors produced larger seedlings with less anthocyanin accumulation than controls. Seedlings of this line also accumulated less anthocyanin on low nitrogen media, indicating a low nitrogen tolerant phenotype; and

seedlings of one of ten G3723 overexpressing lines were more tolerant to desiccation and mannitol (an assay often used to indicate drought tolerance). Two other lines produced larger seedlings than controls on control growth media.

Several of the sequences that fall just outside of the G1274 clade (outside of the large box in Exhibit D), have conserved domains 55%-61% identical to the conserved domain of G1274, including G2517 (overexpressor seedlings were larger and more desiccation tolerant than controls), G194 (one line produced larger seedlings, all four lines tested were more desiccation tolerant than controls), G1758 (several lines of overexpressors had greater tolerance to low nitrogen conditions and chilling during growth than controls), and G179 (one line of G179 overexpressors produced seedlings that were larger and more tolerant to desiccation than controls) also conferred presently claimed improved traits in overexpressing plants.

The remainder of the sequences shown in Exhibit D that lie just outside of the clade have not as yet been shown to confer improved traits similar to those observed in G1274 overexpressors, but these sequences have not been tested extensively or examined in the present study.

6. I hereby declare that all statements made herein are true and that they are based on my own knowledge, information and belief. These statements are made with the knowledge that willful false statements are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of this application or any patent issued from it.

Date: February 13, 2006

A handwritten signature in cursive script, appearing to read "Peter Repetti", written over a horizontal line.

Peter Repetti, Ph.D.  
Senior Scientist  
Mendel Biotechnology, Inc.

PR/jml  
MBI-0054US.Decl.PRepetti.ROA.doc

### Exhibit A. G1274-Related Polypeptide Sequences

#### Sequences derived from *Arabidopsis thaliana*

G1274 (SEQ ID NO: 194)

MNISQNPSPNFTYFSDENFINPFMDNNDFSNLMFFDIDEGGNGLIEEEISSPTSIVSSETFTGESGGSG  
SATTLKKKESTNRGSKESDQTKETGHRVAFRTRSKIDVMDDGFKWRKYGKKSVKNNINKRNYKC  
SSEGCSVKRVERDGDAAAYVITTYEGVHNHESLSNVYYNEMVLSYDHDNWNQHSLRS

G1275

MNDADTNLGSSFSDDTHSVFEPELDLSDEWMDDDLVSAVSGMNQSYGYQTSADVAGALFSGSSSC  
FSHPESPSTKTYVAATATASADNQNKKKKIKGRVAFKTRSEVEVLDDDGFKWRKYGKKMVKNP  
HPRNYYKCSVDGCPVKRVERDRDDPSFVITTYEGSHNHSSMN

G2517

MENVGVGMPFYDLGQTRVYPLLSDFDLSAERYPVGFMDLLGVHRHTPTHTPLMHFPTTPNSSSE  
AVNGDDEEEEDGEEQQHKTKKRKFCTKMSRKQTKKKVPKVSFTRSEVLHLDDGYKWRKYGQKP  
VKDSFPFRNYYRCTTTWCDVKKRVERSFSDPSSVITTYEGQHTHPRPLLIMPEGSSPSNGSASRAHI  
GLPTLPPQLLDYNNQQQAPSSFGTEYNRQEKGINHDDDDHVVKSRTRDLLDGAGLVKDHGLL  
QDVVPSHIKEY

G179

MEDRRCDVLFPCSSSVDPRLTEFHGVDNSAQPTTSSEEKPRSKKKKKEREARYAFQTRSQVDILDDG  
YRWRKYGQKAVKNNPFPRSYKCTEEGCRVKKQVQORQWGDEGVVVTTYQGVHHA  
VDKPSDNFHHILTQMHIFPPFCLKE

G194

MEFTDFSKTsfyypssqsvwdfgdlaaaerhslgmellssqqhqdfatvspHsflLQTSQPQTQTQ  
PSAKLSSSIIQAPPSEQLVTSKVESLCSHLLINPPATPNSSSISSEALNEEKPKTEDNEEEGGEDQ  
QEKSHTKKQLKAKKNNQKRQREARVAFMTKSEVDHLEDGYRWRKYGQKAVKNSPFPRSYRCTT  
ASCNVKKRVERSFRDPSTVVTTYEGQHTHISPLTSRPISTGGFFGSSGAASSLGNGCFGFPIDGSTLISP  
QFQQLVQYHHQQQQELMSCFGGVNEYLNSHANEYGDDNRVKKSRVLVKDNGLLQDVVPSHML  
KEE

G1758

MNYPSPNPSPSTDFTEFFKFDDFDDTFEKIMEEIGREDHSSSPTLSWSSSEKLVA AEITSPLQT  
SLATSPMSFEIGDKDEIKKRKRHKEDPIIHVFKTKSSIDEKVALDDDGYKWRKYGKKPITGSP  
FPRHYHKCSSPDCNVKKKIERDTNPNPDYILTTYEGRHNHPSPSVVYCDSDDFDLNSLNNWS  
FQTANTYSFSHSAPY

#### Sequences derived from *Glycine max*

G3723

MDYYFGNLNPNPYHHS AVVNMASPSSEFMLS DYLVLEDALVVDHHQESWSQSTETESSEKATSS  
DASHGFGDATSNNTNMHIKCQNSGIKGKNAEVSQRITFRTRSQLEVMDDGYKWRKYGKKTVKSSPN  
PRNYYKCSGEGCDVKRVERDRDSDNYVLTTYDGVHNHQTPTAYYSQMPLLSNHDWALHPSA  
NS

G3724 (predicted from polynucleotide SEQ ID NO: 969)

MDFYFGNSPPYPNNYAHNSLNMALSSPEIALSDYLMDDYVDHQDSRSSQSTESSEKATFNDATHG  
FSTGATSKNNNINCKNGINENKGGVGPRIAFRTKSELEIMDDGYKWRKYGKKSVKSSPNLRNYYKC  
SSGGCSVKKRVERDRDDYSYVITTYEGVHNHESPFTTYSPISFVHSDTTFK

G3803

MDYYFGNPNPKPYDNRHSAVVNTESPSSEFMLS DYLVLEDAVDNQESWSQSTETESSEKGNSSDVS  
HGFGDATFSNTNMHIKCENNGIKRKKEEVSQMITFRTRSQLEVMDDGYKWRKYGKKTVKNNPNPR  
NYYKCSGEGCNVKKRVERDRDDSNYVLTTYDGVHNHESPSTAYYSQIPLVHSNHDWPQLHPSANS

**Sequences derived from *Oryza sativa* (japonica cultivar-group)**

G3721

MAASVGLNPEAFFSNSYSYSSSPFMASYTPEFSAAIDANLFSGELDFDCSLPAPAQEYPENENTM  
MRYESEEKMRARVNGRIGFRTRSEVEILDDGFKWRKYGKKAVKNSPNPRNYYRCSTEGCNVKKRV  
ERDREDHRYVITTYDGVHNHASPAALQYAAAAGDYSPPLSSAGSPPAAYFGRRLRCSSEG

G3725 (SEQ ID NO: 970)

MAAAAAGASTPFNFCRHGSHA EYDAVFSGSWMARRPSAAPHGGGASGSGSGSGYGAASYVAPTF  
GAAFRQQHLDLLDYLSDDQGVPAAPPAAVPSASYVTPAPAMAPAEPVVPDAVAAAAGGYPRSVAAA  
AAAVAGEGRDRTTTDKIAFRTRSDDEILDDGYKWRKYGKKSVKNSPNPRNYYRCSTEGCNVKKRV  
ERDKNDPRYVVTMYEGIHNVCPGTVYAAQDAASGRFFVAGISHPDLN

G3726 (SEQ ID NO: 971)

MAAVGAHA AVYHHPVSGLSAPAGDAAYSMSSYFSHGGSSTSSSASSFSAALAAATTPPLPDPSGSQ  
FDISEFFFDDAPPA AVFNGAPTAALPDGAAANATRSAAEAVPAPAPAAVERPRTERIAFRTKSEIEILD  
DGYKWRKYGKKSVKNSPNPRNYYRCSTEGCNVKKRVERDKDDPSYVVTTYEGTHNVSPSTVYY  
ASQDAASGRFFVAGTQPPGSLN

G3729

MSSLYPSLLSLSESPA EYRQVGGGRYAGEDVVD DDDMAAVADAVSSYLSFDMDDVEYYTPEVGF  
HSKQHNPPVAAAPLEAGGREQSRREAAVN LGKMDRGPAPVSGGAATGGVPRSKNGSKIAFKTR  
SEVDVLDDGYRWRKYGKKMVKNSPNPRNYYRCSSEGC RVKKRVERARD DARFVVTTYDGVHNH  
PAPLHLRPQLPPPGYSIAGAPAVVAPHGRLGLEEA EVIALFRGTTATSLLL

G3730

MAASLGLCHETSYAYSYPASNTSSSLCFPPLMADHIVDGGGGGGCSFGEFLELGHSVYSLPLPPPSQ  
PVVVAGGNNDQYGVSSSSAAATTSRIGFRTRSEVEVLDDGFKWRKYGKKAVKSSPNPRNYYRCS  
AAGCGVKKRVERDGD DPRYVVTTYDGVHNHATPGCVGGGGHLPYPTSAAPPWSVPA AAA SPPPA  
HAQAWGAPLHAAAAAHSSSSF

**Sequences derived from *Zea mays***

G3719

MADDYFQFGFDGQEMVGAPAPACGGYDCSAPVFANSSSDAAAAVGNGMSLLSYGVDGDGDGRRP  
MSGPPYGTGGNGGGRPPSSSRIGFRTRSEVDVLDDGFKWRKYGKKTVKSSPNPRNYYRCSAEGCGV  
KKRVERDSDDPRYVVTTYDGVHNHAALGPGAASYLCQPPPPRGATATATVTPFSPPRSASAPLA  
AAPSWSAACDAWEAQLHAAAAAHSSSESY

G3720

MAAVGAHPVLYHHPAPAGDASSMSSYFSHGGSSTTSSSASSFTAALAPTTTALAEHFDISEFLFDDA  
AGAGVAGAPGVFADGAARPVVLVPDAAGGGAIIGAAAAGGAAAASEVPERPRTTTRIAFRTRSEIEIL  
DDGYKWRKYGKKSVKNSPNPRNYYRCSTEGCNVKKRVERDKDDPSYVVTTYEGMHNHVSPSTVY  
YASQDAASGRFFVAGTQPPGSLN

G3722 (predicted from polynucleotide SEQ ID NO: 975)

MHMAISSRSSFAADVLLPATMSYRQPCSGASSYLGSQPAAPFPSAAGAVAQLDVFDCLSSDEGVG  
VPAAVPGAFAPPPPLMPAERVVPDAAAGYSSHTRSAAAVAGEGSRTHRIAFRVRSEDEVLDDGY  
KWRKYGKKSVKNSPNPRNYYRCSTEGCNVKKRVERDRDDPRYVVTMYEGVHNHVSPGTVYYAT  
HDAASGRFFVAGMHQPGH

G3727

MAASLGLNPEAVFTSYTSSPPFMSDYVAASFLPPAVVDSTDFSSELDDLHHHLDYSSPAPTLAGARS  
DRSEKQMRWCEGGGGEKRLGRIGFRTRSEVEILDDGFKWRKYGKKAVKSSPNPRNYYRCSSEGCG  
VKKRVERDRDDPRYVITYDGVHNHASPAAAAIIQYGGGGGFYSPPHSGSPSAASYSGSFVL

G3728

MATSLGLNPEDLFTSYSSSYSSPPFMSDYAASFTPAGGDSTAFSELDDLHHFDYSPAPIVTAAGAG  
AGGGDRNEKMMWCQGGGDERRLRSNGRIGFRTRSQVEILDDGFKWRKYGKKAVKNSPNPRNYYR  
CSSEGCGVKKRVERDRDDPRYVITYDGVHNHASPAAAAIIVPYGSGGGNSGFYSPPHSGSPSATS  
YSGSLAF

G3804 (predicted from polynucleotide SEQ ID NO: 974)

MATSLGLNPEDLFTSYSSSYSSPPFMSDYAASFTPAGGDSTAFSELNLHHFDYSPAPIVTAAGAG  
AGGGDRNEKMMWCQGGGDERRLRSNGRIGFRTRSQVEILDDGFKWRKYGKKAVKNSPNPRNYYR  
CSSEGCGVKKRVERDRDDPRYVITYDGVHNHASPAAAAIIVPYGNGGGNSGFYSPPHSGSPSATS  
YSGSLVF

**Exhibit B. G1274-Related Polypeptide Sequence Conserved Domain Alignments and Assay Results**

In the following alignments of G1274-clade member conserved domains, the first twelve G1274 clade-member sequences from diverse species have at least 74% sequence identity to the conserved domain (amino acid coordinates 111-164) of G1274, SEQ ID NO: 194 and produce larger plants, increased biomass or confer greater tolerance to low water conditions, cold, or low nitrogen conditions, as compared to wild-type control plants, when these sequences are overexpressed in plants. Four sequences that lie just outside the G1274-clade (found outside the large box in Exhibit D) were also able to confer some of the traits that are produced by overexpressing G1274-clade polypeptides in plants. Assays were performed with these sequences overexpressed under the regulatory control of the CaMV 35S promoter unless otherwise indicated.

**G1274 Clade-Member Sequences Conferring Large Size and/or Abiotic Stress Tolerance**

G1274 (*Arabidopsis thaliana*)

Identities = 54/54 (100%)

G1274 : DGFKWRKYGKKSVKNNINKRNYYKCSSEGCSVKKRVERDGDAAAYVITTYEGVH  
DGFKWRKYGKKSVKNNINKRNYYKCSSEGCSVKKRVERDGDAAAYVITTYEGVH

G1274 : DGFKWRKYGKKSVKNNINKRNYYKCSSEGCSVKKRVERDGDAAAYVITTYEGVH

Some lines overexpressing G1274 had larger, broader leaves and greater biomass than controls. Seven of 10 lines of seedlings also had more root growth and root hair than controls under the regulatory control of the emergent leaf primordia-specific AS1 promoter.

G1274 overexpressors were also shown to have greater tolerance to:

- drought (7 of 7 lines tested with G1274 overexpressed under the regulatory control of the 35S promoter)
- desiccation (6 of 30 lines tested with G1274 under the control of the 35S promoter; 4 of 10 lines tested with G1274 under the control of the root-specific ARSK1 promoter, (3 of 10 lines tested with G1274 under the control of the shoot apical meristem-specific STM promoter and 3 of 9 lines tested with G1274 under the control of the vascular tissue-specific SUC2 promoter)
- cold during germination (3 of 10 lines tested with G1274 under the control of the STM promoter)
- cold during growth (12 of 30 lines tested with G1274 under the control of the 35S promoter), and
- nitrogen limitation (6 of 30 lines tested with G1274 under the control of the 35S promoter, 8 of 10 lines tested with G1274 under the control of the epidermal-specific CUT1 promoter, and 4 of 11 lines tested with G1274 under the control of the SUC2 promoter)



G3724 (*Glycine max*; predicted from polynucleotide SEQ ID NO: 969)

Identities = 45/54 (83%)

G1274 : DGFKWRKYGKKS VKNNINKRNYKCSSEGC SVKKRVERDGDAAAYVITTYEGVH  
DG+KWRKYGKKS VK++ N RNYYKCSS GCSVKKRVERD DD +YVITTYEGVH  
G3724 : DGYKWRKYGKKS VKSSPNLRNYYKCSSGGCSVKKRVERDRDDYSYVITTYEGVH

Several 35S::G3724 lines had large rosettes with long, broad leaves relative to controls. Five of 10 lines of seedlings also had more root growth than controls under the regulatory control of the 35S promoter.

35S::G3724 overexpressors were also shown to have greater tolerance to:

- drought (2 of 3 lines tested)
- cold during germination (8 of 10 lines tested)

G3804 (*Zea mays*; predicted from polynucleotide SEQ ID NO: 974)

Identities = 44/54 (81%)

G1274 : DGFKWRKYGKKS VKNNINKRNYKCSSEGC SVKKRVERDGDAAAYVITTYEGVH  
DGFKWRKYGKK+VKN+ N RNYY+CSSEGC VKKRVERD DD YVITTY+GVH  
G3804 : DGFKWRKYGKKAVKNSPNRNYRCSSGCGVKKRVERDRDDPRYVITTYDGVH

35S::G3804 overexpressors were shown to have greater tolerance to:

- drought (3 of 3 lines tested)
- cold during germination (4 of 10 lines tested)

G3803 (*Glycine max*)

Identities = 43/54 (79%)

G1274 : DGFKWRKYGKKS VKNNINKRNYKCSSEGC SVKKRVERDGDAAAYVITTYEGVH  
DG+KWRKYGKK+VKNN N RNYYKCS EGC+VKKRVERD DD+ YV+TTY+GVH  
G3803 : DGYKWRKYGKKT VKNPNRNYKCSGEGCNVKKRVERDRDDSNYVLT TYDGVH

Many of the lines of 35S::G3803 overexpressors developed broad, enlarged leaves and full, bushy rosettes relative to wild-type controls.

35S::G3803 overexpressors were shown to have greater tolerance to:

- desiccation (4 of 20 lines tested)
- cold during germination (3 of 10 lines tested)
- nitrogen limitation (2 of 10 lines tested)

G3727 (*Zea mays*)

Identities = 43/54 (79%)

G1274 : DGFKWRKYGKKS VKNNINKRNYKCSSEGC SVKKRVERDGDAAAYVITTYEGVH  
DGFKWRKYGKK+VK++ N RNYY+CSSEGC VKKRVERD DD YVITTY+GVH  
G3727 : DGFKWRKYGKKAVKSSPNRNYRCSSGCGVKKRVERDRDDPRYVITTYDGVH.

35S::G3727 overexpressors were shown to have greater tolerance to:

- nitrogen limitation (3 of 10 lines tested)

G3720 (*Zea mays*)

Identities = 42/54 (77%)

G1274: DGFKWRKYGKKS VKNNINKRNYKCSSEGCSVKRVERDGDAAAYVITTYEGVH  
DG+KWRKYGKKS VKN+ N RNYY+CS+EGC+VKRVERD DD +YV+TTYEG+H  
G3720: DGYKWRKYGKKS VKNSPNPRNYRCSTEGCNVKRVERDKDDPSYVVTTYEGMH

Three lines of 35S::G3720 plants were very bushy with broad, flat, wavy leaves.

G3720 overexpressors were shown to have greater tolerance to:

- desiccation (1 of 10 lines tested), and
- nitrogen limitation (3 of 10 lines tested)

G3726 (*Oryza sativa*; SEQ ID NO: 971)

Identities = 42/54 (77%)

G1274: DGFKWRKYGKKS VKNNINKRNYKCSSEGCSVKRVERDGDAAAYVITTYEGVH  
DG+KWRKYGKKS VKN+ N RNYY+CS+EGC+VKRVERD DD +YV+TTYEG H  
G3726: DGYKWRKYGKKS VKNSPNPRNYRCSTEGCNVKRVERDKDDPSYVVTTYEGTH

Seedlings of four lines were larger or had more root growth than controls. At least one line of 35S::G3726 overexpressors was larger than controls at the rosette stage, several other lines had flat, broad leaves relative to controls. 35S::G3726 overexpressors were also shown to have greater tolerance to:

- drought (2 of 3 lines tested)
- cold during germination (3 of 10 lines tested), and
- cold during growth (3 of 10 lines tested)

G3722 (*Zea mays*; predicted from polynucleotide SEQ ID NO: 975)

Identities = 42/54 (77%)

G1274: DGFKWRKYGKKS VKNNINKRNYKCSSEGCSVKRVERDGDAAAYVITTYEGVH  
DG+KWRKYGKKS VKN+ N RNYY+CS+EGC+VKRVERD DD YV+T YEGVH  
G3722: DGYKWRKYGKKS VKNSPNPRNYRCSTEGCNVKRVERDRDDPRYVVTTYEGVH

35S::G3722 overexpressors were shown to have greater tolerance to:

- desiccation (2 of 10 lines tested), and
- nitrogen limitation (6 of 10 lines tested)

G3721 (*Oryza sativa*)

Identities = 42/54 (77%)

G1274: DGFKWRKYGKKS VKNNINKRNYKCSSEGCSVKRVERDGDAAAYVITTYEGVH  
DGFKWRKYGKK+VKN+ N RNYY+CS+EGC+VKRVERD +D YVITTY+GVH  
G3721: DGFKWRKYGKKAVKNSPNPRNYRCSTEGCNVKRVERDREDHRYVITTYDGVH

Germinating seedlings of two lines were larger than controls. At least one line of 35S::G3721 overexpressors had broad leaves relative to controls.

35S::G3721 overexpressors were also shown to have greater tolerance to:

- drought (3 of 3 lines tested), and
- cold during germination (10 of 10 lines tested)

G1275 (*Arabidopsis thaliana*)

Identities = 41/54 (75%)

G1274 : DGFKWRKYGKKS VKNNINKRNYYKCSSEGC SVKKRVERDGDAAAYVITTYEGVH  
DGFKWRKYGKK VKN+ + RNYYKCS +GC VKKRVERD DD ++VITTYEG H  
G1275 : DGFKWRKYGKKMVKN SPHPRNYYKCSVDGCPVKKRVERDRDDPSFVITTYEGSH

Seedlings of several 35S::G1275 lines were larger than controls. At least one 35S::G1275 line had broader leaves and a larger, fuller rosette than controls. One CUT1::G1275 line had a larger rosette size with large, flat leaves relative to controls.

G1275 overexpressors were also shown to have greater tolerance to:

- drought (2 of 3 lines tested with G1275 overexpressed under the regulatory control of the CUT1 promoter, and 2 of 3 lines tested with G1275 overexpressed under the control of the stress-inducible RD29A promoter)
- cold during germination (4 of 10 lines tested with G1275 overexpressed under the control of the CUT1 promoter)
- cold during growth (4/10 lines more tolerant, 2/10 lines more sensitive with G1275 overexpressed under the control of the 35S promoter), and
- nitrogen limitation (4 of 10 lines tested with G1275 overexpressed under the control of the CUT1 promoter, 5 of 10 lines tested with G1275 overexpressed under the control of the STM promoter, and 10 of 10 lines tested with G1275 overexpressed under the control of the SUC2 promoter)

G3729 (*Oryza sativa*)

Identities = 40/54 (74%)

G1274 : DGFKWRKYGKKS VKNNINKRNYYKCSSEGC SVKKRVERDGDAAAYVITTYEGVH  
DG++WRKYGKK VKN+ N RNYY+CSSEGC VKKRVER DDA +V+TTY+GVH  
G3729 : DGYRWRKYGKKMVKN SPNPRNYYRCSEGC RVKKRVERARDARFVVTTYDGVH

Three lines of 35S::G3729 overexpressors had flat, enlarged broad leaves.

35S::G3729 plants were also shown to have greater tolerance to:

- cold during germination (5 of 12 lines tested)
- nitrogen limitation (11 of 12 lines tested)

G3725 (*Oryza sativa*; SEQ ID NO: 970)

Identities = 40/54 (74%)

G1274 : DGFKWRKYGKKS VKNNINKRNYYKCSSEGC SVKKRVERDGDAAAYVITTYEGVH  
DG+KWRKYGKKS VKN+ N RNYY+CS+EGC+VKKRVERD +D YV+T YEG+H  
G3725 : DGYKWRKYGKKS VKNSPNPRNYYRCSTEGCNVKKRVERDKNDPRYVVTMYEGIH

Most of the 35S::G3725 lines produced seedlings with more vigorous root growth than controls.

- One line of 35S::G3725 overexpressors had larger seedlings in cold conditions than controls.

### Sequences Near to the G1274 Clade that Conferred Large Size and/or Stress Tolerance

#### G2517 (*Arabidopsis thaliana*)

Identities = 33/54 (61%)

G1274: DGFKWRKYGKKS VKNNINKRNYKCSSEGCSVKKRVERDGDDAAYVITTYEGVH  
DG+KWRKYG+K VK++ RNY+ C++ C VKKRVER D + VITTYEG H  
G2517: DGYKWRKYGQKPVKDSFPFRNYRCTTTWC DVKKRVERSFSDPSSVITTYEGQH

G2517 lies just outside of the phylogenetically defined G1274 clade (Exhibit D). This gene was included in the present study to test the boundary of the G1274 clade, and determine the range of effects caused by genes that are just outside the G1274 clade.

35S::G2517 overexpressors produced seedlings that were often larger than controls. Two lines overexpressing G2517 had larger rosettes with broader, rounder leaves than controls.

35S::G2517 overexpressors were also shown to have greater tolerance to:

- desiccation (6 of 10 lines tested)

#### G194 (*Arabidopsis thaliana*)

Identities = 31/54 (57%)

G1274: DGFKWRKYGKKS VKNNINKRNYKCSSEGCSVKKRVERDGDDAAYVITTYEGVH  
DG++WRKYG+K+VKN+ R+YY+C++ C+VKKRVER D + V+TTYEG H  
G194: DGYRWRKYGQKAVKNSPFRSYRCTTASCNVKKRVERSFDPSTVVTTYEGQH

G194 lies just outside of the phylogenetically defined G1274 clade (Exhibit D). G194 is very closely related to G2517, which produced similar stress tolerance effects as G194 when overexpressed.

One line of 35S::G194 overexpressors produced larger seedlings than controls. Two lines were bushy in appearance and had more rosette leaves with profuse branching relative to controls.

35S::G194 overexpressors were also shown to have greater tolerance to:

- desiccation (4 of 4 lines tested)

#### G1758 (*Arabidopsis thaliana*)

Identities = 30/54 (55%)

G1274: DGFKWRKYGKKS VKNNINKRNYKCSSEGCSVKKRVERDGDDAAYVITTYEGVH  
DG+KWRKYGKK + + R+Y+KCSS C+VKK++ERD ++ Y++TTYEG H  
G1758: DGYKWRKYGKKPITGSPFRHYHKCSSPDCNVKKKIERDTNPNPDYILTTYEGRH

G1758 lies just outside of the phylogenetically defined G1274 clade (Exhibit D).

35S::G1758 overexpressors were shown to have greater tolerance to:

- cold during growth (3 of 10 lines tested)
- nitrogen limitation (2 of 10 lines tested)

The following sequences have been overexpressed in plants and have not yet produced larger plants or increased abiotic stress tolerance in a significant number of lines. Significance may yet be achieved by generating and studying a greater number of overexpressing lines. Nonetheless, each of these sequences did demonstrate the same improved traits that were conferred by G1274 clade members in some of the lines tested.

#### **G1274 Clade members**

##### G3728 (*Zea mays*)

Identities = 44/54 (81%)

G1274 : DGFKWRKYGKKS VKNNINKRNYYKCSSEGCSVKKRVERDGDAAAYVITTYEGVH  
DGFKWRKYGKK+VKN+ N RNYY+CSSEGCVKKRVERD DD YVITTY+GVH  
G3728 : DGFKWRKYGKKA VKNSPNRNYYRCSSEGCGVKKRVERDRDDPRYVITTYDGVH

Some 35S::G3728 seedlings of one line were slightly more tolerant to desiccation than controls in a plate-based assay. One line was larger than controls at the rosette stage. Seedlings of 4 of 60 35S::G3728 lines examined also appeared to be larger than controls.

##### G3719 (*Zea mays*)

Identities = 41/54 (75%)

G1274 : DGFKWRKYGKKS VKNNINKRNYYKCSSEGCSVKKRVERDGDAAAYVITTYEGVH  
DGFKWRKYGKK+VK++ N RNYY+CS+EGC VKKRVERD DD YV+TTY+GVH  
G3719 : DGFKWRKYGKKT VKSSPNRNYYRCSAEGCGVKKRVERDSDDPRYVVTTYDGVH

Only four lines of overexpressors were tested in physiological assays. One line accumulated less anthocyanin than controls on low nitrogen media, indicating a low nitrogen tolerant phenotype.

##### G3730 (*Oryza sativa*)

Identities = 41/54 (75%)

G1274 : DGFKWRKYGKKS VKNNINKRNYYKCSSEGCSVKKRVERDGDAAAYVITTYEGVH  
DGFKWRKYGKK+VK++ N RNYY+CS+ GC VKKRVERDGD DD YV+TTY+GVH  
G3730 : DGFKWRKYGKKA VKSSPNRNYYRCSAAGCGVKKRVERDGDPRYVVTTYDGVH

One line of 35S::G3730 overexpressors had broad, round leaves relative to controls.

35S::G3730 plants were also shown to have greater tolerance to:

- cold during germination (one of four 35S::G3730 lines produced larger seedlings with less anthocyanin accumulation than controls.
- nitrogen limitation (seedlings of the cold tolerant line also accumulated less anthocyanin on low nitrogen media, indicating a low nitrogen tolerant phenotype).

##### G3723 (*Glycine max*)

Identities = 41/54 (75%)

G1274 : DGFKWRKYGKKS VKNNINKRNYYKCSSEGCSVKKRVERDGDAAAYVITTYEGVH  
DG+KWRKYGKK+VK++ N RNYYKCS EGC VKKRVERD DD+ YV+TTY+GVH  
G3723 : DGYKWRKYGKKT VKSSPNRNYYKCSGEGCDVKKRVERDRDDSNYVLT TYDGVH

One line of 35S::G3723 overexpressors had long, broad, leaves relative to controls. One line also produced seedlings that were slightly larger than controls on control growth media. Another line had larger roots than controls.

35S::G3723 plants were also shown to have greater tolerance to:

- desiccation (1 of 10 lines tested; this same line was also more tolerant to mannitol in a germination assay, an assay often used to indicate drought tolerance).

### **Sequences Near the G1274 Clade**

#### G179 (*Arabidopsis thaliana*)

Identities = 33/54 (61%)

G1274:    DGFKWRKYGKKS VKNNINKRNYKCSSEGCSVKKRVERDGDDAAYVITTYEGVH  
          DG++WRKYG+K+VKNN    R+YYKC+ EGC VKK+V+R    D    V+TTY+GVH  
G179:    DGYRWRKYGQKAVKNNPFPRSYKCTEEGCRVKKQVQRQWGDEGVVTTYQGVH

G179 lies just outside of the phylogenetically defined G1274 clade (Exhibit D). One of ten 35S::G179 plants produced seedlings that were larger than control seedlings on control growth media. This same line also produced seedlings that were more tolerant to desiccation than controls in a plate-based germination assay.



### Exhibit C. Assay Methods

Plate-based physiological assays representing a variety of stress related conditions were used to identify stress-tolerant lines. In addition, nitrogen limitation studies were performed to find genes that allowed more plant growth upon deprivation of nitrogen. Nitrogen is a major nutrient affecting plant growth and development that ultimately impacts yield and stress tolerance. These assays monitored primarily root but also rosette growth on nitrogen deficient media. In all higher plants, inorganic nitrogen is first assimilated into glutamate, glutamine, aspartate and asparagine, the four amino acids used to transport assimilated nitrogen from sources (e.g. leaves) to sinks (e.g. developing seeds). This process is regulated by light, as well as by C/N metabolic status of the plant. We used a C/N sensing assay to look for alterations in the mechanisms plants use to sense internal levels of carbon and nitrogen metabolites which could activate signal transduction cascades that regulate the transcription of N-assimilatory genes. To determine whether these mechanisms are altered, we exploited the observation that wild-type plants grown on media containing high levels of sucrose (3%) without a nitrogen source accumulate high levels of anthocyanins. This sucrose induced anthocyanin accumulation can be relieved by the addition of either inorganic or organic nitrogen. We used glutamine as a nitrogen source since it also serves as a compound used to transport N in plants.

Prior to plating, seed for all experiments were surface sterilized in the following manner: (1) 5 minute incubation with mixing in 70 % ethanol, (2) 20 minute incubation with mixing in 30% bleach, 0.01% triton-X 100, (3) 5X rinses with sterile water, (4) Seeds were re-suspended in 0.1% sterile agarose and stratified at 4 °C for 3-4 days.

#### Germination assays

Cold germination assays were conducted at 8 °C. -N media was basal media minus nitrogen plus 3% sucrose. -N/+Gln media was basal media minus nitrogen plus 3% sucrose and 1 mM glutamine.

All germination assays followed modifications of the same basic protocol. Sterile seeds were sown on the conditional media that had a basal composition of 80% MS + Vitamins. Plates were incubated at 22 °C under 24-hour light ( $120\text{-}130 \mu\text{E m}^{-2} \text{s}^{-1}$ ) in a growth chamber. Evaluation of germination and seedling vigor was performed 5 days after planting. For assessment of root development, seedlings germinated on 80% MS + Vitamins + 1% sucrose were transferred to square plates at 7 days. Evaluation was performed 5 days after transfer following growth in a vertical position. Qualitative differences were recorded including lateral and primary root length, root hair number and length, and overall growth.

#### Growth assays

Severe desiccation assays were conducted for 5 days followed by recovery at 22 °C. Growth in cold or chilling conditions were conducted at (8 °C). Root development analysis consisted of visual assessment of lateral and primary roots, root hairs and overall growth.

For the nitrogen limitation assays, all components of MS medium remained constant except N was reduced to 20mg/L of  $\text{NH}_4\text{NO}_3$ . Note that 80% MS has 1.32 g/L  $\text{NH}_4\text{NO}_3$  and 1.52 g/L  $\text{KNO}_3$ .

Experiments are performed with the *Arabidopsis thaliana* ecotype Columbia (col-0). Assays were usually performed on non-selected segregating T2 populations in order to avoid the extra stress of selection. Control plants for assays on lines containing direct promoter-fusion constructs were wild-type or Col-0 plants transformed an empty transformation vector (pMEN65). Controls for 2-component lines (generated by supertransformation) were wild type or the background promoter-driver lines (i.e. promoter::LexA-GAL4TA lines), into which the supertransformations were initially performed.

All assays were performed in tissue culture. Growing the plants under controlled temperature and humidity on sterile medium produces uniform plant material that has not been exposed to additional stresses (such as water stress) which could cause variability in the results obtained. All assays were designed to detect plants that were more tolerant or less tolerant to the particular stress condition and were developed with reference to the following publications: Jang et al. (1997) *Plant Cell* 9: 5-19, Smeekens (1998) *Curr. Opin. Plant Biol.* 1: 230-234, Liu and Zhu (1997) *Proc. Natl. Acad. Sci. USA* 94: 14960-14964, Saleki et al. (1993) *Plant Physiol.* 101: 839-845, Xu et al. (1996) *Plant Physiol.* 110: 249-257, Zhu et al. (1998) *Plant Cell* 10: 1181-1191, Alia et al. (1998) *Plant J.* 16: 155-161, Xin and Browse (1998) *Proc. Natl. Acad. Sci. USA* 95: 7799-7804, and Leon-Kloosterziel et al. (1996) *Plant Physiol.* 110: 233-240. Where possible, assay conditions were originally tested in a blind experiment with controls that had phenotypes related to the condition tested.

For chilling (8 °C) growth assays, seeds were germinated and grown for 7 days on MS + Vitamins + 1% sucrose at 22 °C and then are transferred to chilling conditions (8 °C) and evaluated at 10 days and 17 days.

For severe desiccation (water deprivation) assays, seedlings were grown for 14 days on MS+ Vitamins + 1% Sucrose at 22 °C. Plates were opened in the sterile hood for 3 hr for hardening and then seedlings were removed from the media and dried for two hours in the hood. After this time they were transferred back to plates and incubated at 22 °C for recovery. Plants were evaluated after 5 days.

#### Soil Drought Assays

The soil drought assay (performed in clay pots) is based on that described by Haake et al. (2002).

Seeds are sterilized by a 2 minute ethanol treatment followed by 20 minutes in 30% bleach / 0.01% Tween and five washes in distilled water. Seeds are sown to MS agar in 0.1% agarose and stratified for 3 days at 4 °C, before transfer to growth cabinets with a temperature of 22 °C. After 7 days of growth on selection plates, seedlings were transplanted to 3.5 inch diameter clay pots containing 80g of a 50:50 mix of vermiculite:perlite topped with 80g of ProMix. Typically, each pot contained 14 seedlings, and plants of the transgenic line being tested were in separate pots to the wild-type controls. Pots containing the transgenic



line versus control pots were interspersed in the growth room, maintained under 24-hour light conditions ( $18 - 23\text{ }^{\circ}\text{C}$ , and  $90 - 100\text{ }\mu\text{E m}^{-2}\text{ s}^{-1}$ ) and watered for a period of 14 days. Water was then withheld and pots were placed on absorbent diaper paper for a period of 8-10 days to apply a drought treatment. After this period, a visual qualitative "drought score" from 0-6 was assigned to record the extent of visible drought stress symptoms. A score of "6" corresponded to no visible symptoms whereas a score of "0" corresponded to extreme wilting and the leaves having a "crispy" texture. At the end of the drought period, pots were re-watered and scored after 5-6 days; the number of surviving plants in each pot was counted, and the proportion of the total plants in the pot that survived was calculated.

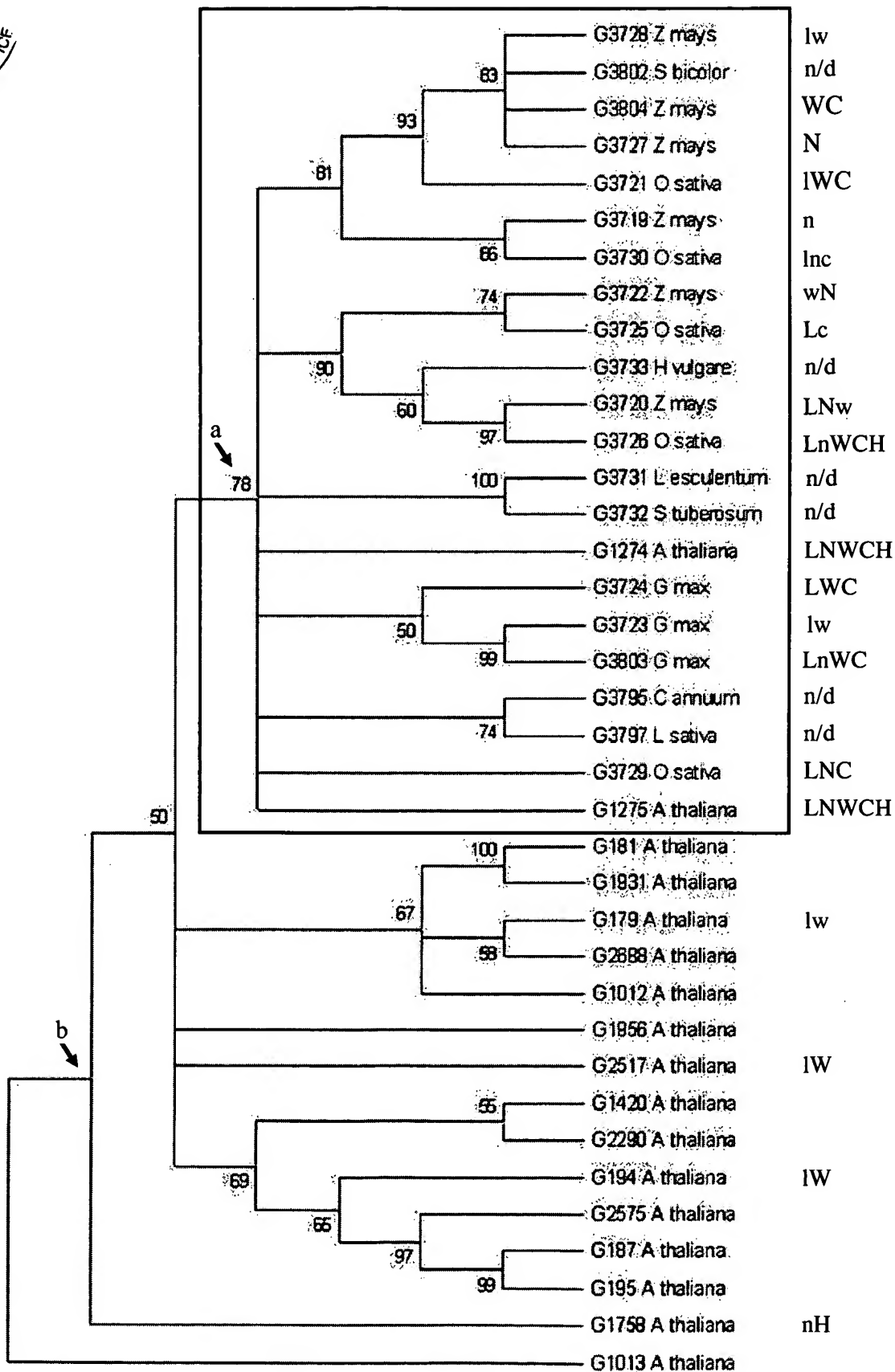
Analysis of results. In a given experiment, we typically compared six or more pots of a transgenic line with six or more pots of the appropriate control. The mean drought score and mean proportion of plants surviving (survival rate) were calculated for both the transgenic line and the wild-type pots. In each case a  $p$ -value\* was calculated, which indicated the significance of the difference between the two mean values. The results for each transgenic line across each planting for a particular project were then presented in a results table.

Calculation of p-values. Survival was analyzed with a logistic regression to account for the fact that the random variable is a proportion between 0 and 1. The reported  $p$ -value is the significance of the experimental proportion contrasted to the control, based upon regressing the logit-transformed data.

Drought score, being an ordered factor with no real numeric meaning, was analyzed with a non-parametric test between the experimental and control groups. The  $p$ -value was calculated with a MannWhitney rank-sum test.



# Exhibit D. G1274 Clade Phylogenetic Analysis



Sequences shown to confer improved traits similar to those conferred in G1274 overexpressors are found in the phylogenetic tree in this Exhibit. These traits, which were observed relative to control plants, may be identified by the following symbols:

- L large plants, large plant organs (e.g., leaves) or increased biomass
- N low N tolerant
- W water deprivation tolerant as determined in drought and/or desiccation assays
- C cold germination tolerant
- H chilling during growth tolerant
- n/d transformation and/or assay not yet performed

Lower case letters indicate a trait that was observed in fewer than three lines, or, in the case of G3728 and increased biomass, the increase was slight relative to controls. Thus far, only *Arabidopsis*, soy, rice and maize sequences have been overexpressed in plants that were examined for these traits. Only four of the *Arabidopsis* sequences below the large box, G179, G2517, G194, and G1758 were examined in the present study.

Each of the G1274-related sequences were included in the present study because of their phylogenetic relatedness to the G1274 sequence. Each of these polypeptide sequences descend from common ancestral sequences. One such ancestral sequence is represented by the node at arrow "a", from which the G1274 clade descends; the G1274 clade is encompassed by the large box overlaying the phylogenetic tree. The G1274 clade polypeptides comprise conserved domains, indicated by the underlined residues in Exhibit A, that are at least 74% identical to the conserved domain of G1274, amino acid coordinates 111-164.

Sequences that lie just outside of the clade (derived from a somewhat more distant ancestral sequence at arrow "b") also comprise conserved domains similar to that found in G1274, but these are less similar to G1274 than the sequences within the clade. For example, G2517, G194 and G1758 G179 have conserved domains, indicated by the respective underlined residues in Exhibit A, that are 55% to 61% identical to the conserved domain of G1274.

22. R. Cooke, M. S. Crowder, C. H. Wendt, V. A. Barnett, D. D. Thomas, in *Contractile Mechanisms in Muscle*, G. Pollack and H. Sugi, Eds. (Plenum, New York, 1984), pp. 413-423; H. E. Huxley and M. Kress, *J. Muscle Res. Cell Motil.* 6, 153 (1985).
23. K. Wakabayashi et al., *Science* 258, 443 (1992); S. Highsmith and D. Eden, *Biochemistry* 32, 2455 (1993).
24. K. Burton, *J. Muscle Res. Cell Motil.* 13, 590 (1992); I. Rayment et al., *Science* 261, 58 (1993); R. R. Schröder et al., *Nature* 364, 171 (1993).
25. T. Q. P. Uyeda, S. J. Kron, J. A. Spudich, *J. Mol. Biol.* 214, 699 (1990); Y. Harada, K. Sakurada, T. Aoki, D. D. Thomas, T. Yanagida, *ibid.* 216, 49 (1990). We centrifuged the wild-type myosin and My $\Delta$ RLCBS solutions in the presence of 2 mM MgATP and actin (0.1 mg/ml) to remove denatured molecules and then diluted them to 150 and 300  $\mu$ g/ml, respectively, in the motility assay buffer [25 mM imidazole (pH 7.4), 25 mM KCl, 4 mM MgCl<sub>2</sub>, 1 mM EGTA, and 10 mM DTT] supplemented with 200 mM NaCl. These solutions were assayed in the sliding filament motility system at 30°C.
26. T. T. Egelhoff, S. B. Brown, J. A. Spudich, *J. Cell Biol.* 112, 677 (1991); T. T. Egelhoff, R. J. Lee, J. A. Spudich, *Cell* 75, 363 (1993). All three threonine residues on the Dictyostelium myosin II heavy chain capable of being phosphorylated were substituted with alanines. Cells expressing this mutant myosin (My3X ALA) can divide in suspension culture and form fruiting bodies, albeit inefficiently.
27. T. Q. P. Uyeda and J. A. Spudich, unpublished data. The tail part of the My $\Delta$ RLCBS *mhcA* gene was replaced with that of My3X ALA. The resultant *mhcA* gene, My $\Delta$ RLCBS/3X ALA, was expressed in HS1 cells. Unlike the parent HS1 cells, the transformants could divide in a suspension culture of HL-5 medium supplemented with killed bacteria. However, unlike wild-type cells and My $\Delta$ RLCBS cells, the My $\Delta$ RLCBS/3X ALA cells grew poorly in a suspension culture of regular HL-5. These growth properties are commonly seen with cells expressing the My3X ALA myosin and therefore seem to be a consequence of the loss of regulation by heavy chain phosphorylation. The developmental phenotype of the My $\Delta$ RLCBS/3X ALA cells was different from that of the My3X ALA cells in that the My $\Delta$ RLCBS/3X ALA cells could not proceed beyond the mound stage.
28. K. M. Ruppel, T. Q. P. Uyeda, J. A. Spudich, in preparation.
29. Assay conditions were 10 mM imidazole (pH 7.4), 0.6 M KCl, 5 mM CaCl<sub>2</sub>, 1 mM DTT, myosin (0.1 mg/ml), and 3 mM [ $\gamma$ -<sup>32</sup>P]ATP (~1 mCi/mmol) for the Ca<sup>2+</sup>-ATPase activity and 25 mM imidazole (pH 7.4), 25 mM KCl, 4 mM MgCl<sub>2</sub>, 1 mM DTT, myosin (0.1 mg/ml), and 3 mM [ $\gamma$ -<sup>32</sup>P]ATP in the presence or absence of rabbit skeletal muscle F-actin (1 mg/ml) for the Mg<sup>2+</sup>-ATPase activity. Liberated inorganic phosphate (P<sub>i</sub>) was quantitated as described [M. Ikebe and D. J. Hartshorne, *Biochemistry* 24, 2380 (1985)].
30. T. Maita et al., *J. Biochem.* 110, 75 (1991).
31. H. M. Warrick, A. DeLozanne, L. A. Leinward, J. A. Spudich, *Proc. Natl. Acad. Sci. U.S.A.* 83, 9433 (1986).
32. M. Sussman, *Methods Cell Biol.* 28, 9 (1987).
33. G. Peltz, J. A. Spudich, P. Parham, *J. Cell Biol.* 100, 1016 (1985).
34. We thank I. Rayment for sharing information before publication and K. M. Ruppel and H. M. Warrick for discussions and help with the preparation of the manuscript. Supported by a postdoctoral fellowship from the American Heart Association, California Affiliate (to T.Q.P.U.), and by grant GM46551 from the NIH (to J.A.S.).

18 June 1993; accepted 19 October 1993

## Isolation of *ORC6*, a Component of the Yeast Origin Recognition Complex by a One-Hybrid System

Joachim J. Li\*† and Ira Herskowitz

Here a method is described to identify genes encoding proteins that recognize a specific DNA sequence. A bank of random protein segments tagged with a transcriptional activation domain is screened for proteins that can activate a reporter gene containing the sequence in its promoter. This strategy was used to identify an essential protein that interacts in vivo with the yeast origin of DNA replication. Matches between its predicted amino acid sequence and peptide sequence obtained from the 50-kilodalton subunit of the yeast origin recognition complex (ORC) established that the gene isolated here, *ORC6*, encodes this subunit. These observations provide evidence that ORC recognizes yeast replication origins in vivo.

The replication of DNA in eukaryotic cells is tightly controlled and coordinated with other events in the cell division cycle. This control is thought to be exerted primarily at the initiation of DNA replication. Replication initiation depends on

the completion of earlier events in the cell cycle that commit the cell to a new round of cell division, and reinitiation is prevented until later events are completed, particularly mitosis.

Eukaryotic chromosomal replication initiates at multiple sites in the genome and proceeds bidirectionally. The position of these sites is believed to be specified by DNA elements called origins of replication. Much of our knowledge about the initiation of bidirectional replication comes from prokaryotic and viral systems, most notably the

replication systems of *Escherichia coli*, phage  $\lambda$ , and SV40. In these systems, replication origins have been identified, and in vitro systems are available to dissect the initiation reaction (1). Studies on the initiation of eukaryotic DNA replication, however, have been hampered by difficulty in the identification of origin sequences. Putative origins have been isolated in a number of eukaryotic systems (2), but proof of origin function has remained elusive, and the definition of these elements at the nucleotide level has proven frustrating.

Only in the yeast *Saccharomyces cerevisiae* have eukaryotic origin sequences been clearly identified [reviewed in (3)]. Yeast origins were first detected as DNA elements that allow plasmids to be maintained autonomously in yeast cells and were called autonomous replicating sequences (ARSs). ARSs act as replication origins on plasmids and, in many cases, behave as origins in their native chromosomal location [reviewed in (4)]. ARSs have a bipartite structure. Domain A is primarily composed of a degenerate 11-bp ARS consensus sequence (ACS), 5'-(T/A)-TTTA(T/C)(A/G)TTT(T/A)-3' found in virtually all ARSs (5). Domain B, which is approximately 100 bp in size and positioned 3' to the T-rich strand of domain A, exhibits little sequence similarity among ARSs and appears to be organized from multiple partially redundant sequence elements (6). Because the ACS is the only sequence motif common to all known ARSs, and because single point mutations in this sequence can abolish ARS activity (5), proteins that specifically recognize the ACS are prime candidates for proteins that initiate DNA replication.

In order to identify potential yeast initiators, we developed a genetic strategy (Fig. 1), the one-hybrid system, to find proteins that recognize a target sequence of interest. This strategy was derived from the two-hybrid system for detecting protein-protein interactions (7). The one-hybrid system has two basic components: (i) a hybrid expression library, constructed by fusing a transcriptional activation domain to random protein segments, and (ii) a reporter gene containing a binding site of interest within its promoter region. Hybrid proteins that recognize this site are expected to induce expression of the reporter gene because of their dual ability to bind the promoter region and activate transcription (8). This association may be indirect because hybrids that interact with endogenous proteins already occupying the binding site can also activate transcription (7). Nevertheless, as long as the association is sequence-specific one may expect the protein incorporated in the hybrid to be functionally relevant.

We have used this method to look for proteins from the yeast *Saccharomyces cerevisiae* that recognize the ACS of yeast origins

Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, CA 94143-0448.

\*To whom correspondence should be addressed.

†Present address: Department of Microbiology and Immunology, University of California, San Francisco, CA 94143-0414.

of DNA replication. The protein component of this screen was provided by a set of three complementary yeast hybrid expression libraries, YL1-3, which contain random yeast protein segments fused to the GAL4 transcriptional activation domain (GAL4<sup>AD</sup>) (9). The reporter gene for our screen contained four direct repeats of the ACS in its promoter region and was integrated into the yeast strain GGY1 to form JLY363(ACS<sup>WT</sup>) (Fig. 2) (10). To determine the dependence of *lacZ* induction on the ACS, we constructed in parallel JLY365(ACS<sup>MUTANT</sup>), which harbors a reporter gene carrying four copies of a non-functional ACS that is multiply mutated (Fig. 2) (10).

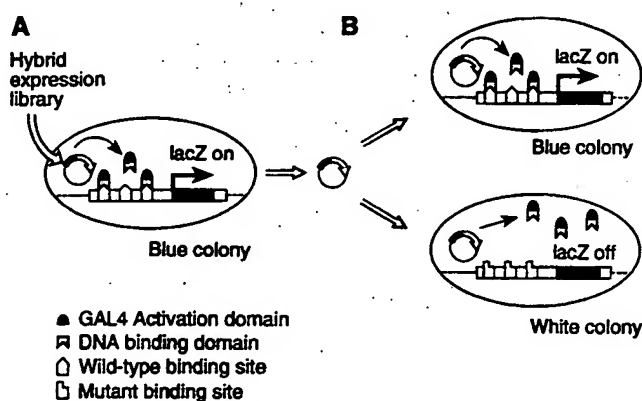
With the strategy depicted (Fig. 1), we isolated nine plasmids that induced greater *lacZ* activity in JLY363(ACS<sup>WT</sup>) than in JLY365(ACS<sup>MUTANT</sup>) from a screen of 1.2 million YL1-3 transformants (11). Many of the plasmids that induced *lacZ* activity on initial screening of the library in JLY363(ACS<sup>WT</sup>) failed to exhibit a dependence on the ACS when introduced into JLY365(ACS<sup>MUTANT</sup>) (Fig. 3). Restriction analysis of these plasmids showed that the nine isolates represented five genomic clones, which we initially labeled AAP1-5 for ACS associated protein. AAP1 was isolated four times, AAP5 twice, and the others only once.

To examine the sequence specificity of

*lacZ* induction with greater resolution, reporter constructs containing direct repeats of four ACS point mutants were each integrated into GGY1 to generate the set of reporter strains tabulated in Fig. 2 (10). The five AAP clones were individually examined in these strains for the ability to induce *lacZ* expression (Fig. 4). AAP1 displayed a correspondence between the induction of this set of reporter genes and the ARS function (12) of their ACS (Fig. 4, top row). The AAP5 hybrid exhibited a slightly weaker correlation, and the remaining clones showed poor correlation (13). These findings suggest that AAP1, and possibly AAP5, encodes a protein that recognizes the ACS in a sequence-specific manner. Constructs with deletions in the AAP1 coding sequence (14) were unable to induce *lacZ* expression (Fig. 4), indicating that recognition of the ACS resided in the protein segment fused to GAL4.

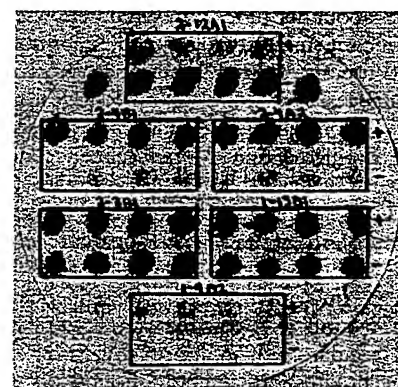
The genomic segments fused to the GAL4<sup>AD</sup> in AAP1-5 were sequenced (15) to determine the extent of the hybrid proteins that were made. AAP1 and AAP5 had sizable protein coding sequences of 301 and 123 amino acids, respectively, fused in frame with the GAL4<sup>AD</sup>. In principle, these segments are large enough to direct the hybrid protein to the promoter of the reporter gene. AAP2-4 encoded hybrid proteins with only short peptide extensions (10, 22, and 38 amino acids, respectively) fused to the GAL4<sup>AD</sup>, suggesting that these

**Fig. 1.** Schematic of one-hybrid system screen for identifying proteins that can recognize a binding site of interest. (A) An expression library of hybrid proteins is transformed into a reporter strain. The hybrids contain protein coding sequences fused to the end of a constitutively expressed GAL4 activation domain. The reporter strain contains a UAS-less *lacZ* reporter gene with multiple copies of the binding site in its promoter region and a low basal transcriptional activity. Hybrid proteins that recognize the binding site act as transcriptional activators of the reporter gene and turn the cell blue in a  $\beta$ -galactosidase assay. (B) Reporter genes containing either wild-type (top) or mutant (bottom) binding sites in their promoter regions are used to test the sequence specificity of the *lacZ* induction observed in (A). Recovered plasmids are introduced into strains carrying one or the other reporter gene, and *lacZ* expression is compared.



| Mutant designation | Oligonucleotide pair   | Two repeats   |                 | Four repeats  |                 |
|--------------------|--|---------------|-----------------|---------------|-----------------|
|                    |  | Reporter gene | Reporter strain | Reporter gene | Reporter strain |
| WT                 | 5'GATCgaattCAGATTTTATGTTTA 3'<br>3'gcttaaGTCTAAAATACAAATCTAG 5'  | pJL623        | JLY360          | pJL625        | JLY363          |
| Multiple           | 5'GATCgaattCAGATATATTTCTATA 3'<br>3'gcttaaGTCTTATATAGATATCTAG 5' | pJL624        | JLY361          | pJL626        | JLY365          |
| A863T              | 5'GATCgaattCAGATTTTATGTTTA 3'<br>3'gcttaaGTCTAAAAACAAATCTAG 5'   | pJL696        | JLY429          | —             | —               |
| T859A              | 5'GATCgaattCAGATTTTATGTATA 3'<br>3'gcttaaGTCTAAAATACATATCTAG 5'  | pJL697        | JLY431          | —             | —               |
| T862C              | 5'GATCgaattCAGATTTTACGTTTA 3'<br>3'gcttaaGTCTAAAATGCAATCTAG 5'   | JL698         | JLY433          | —             | —               |
| T867G              | 5'GATCgaattCAGATTTTATGTTTA 3'<br>3'gcttaaGTCTCAAATACAAATCTAG 5'  | pJL699        | JLY435          | —             | —               |

**Fig. 2.** Reporter genes and strains generated from mutant domain A sequences. Complementary pairs of oligonucleotides representing wild-type and mutant domain A sequences from ARS1 are shown. Reporter gene constructs were generated by inserting two or four direct repeats of these sequences into pBgl-lacZ. Integration of each construct into GGY1 produced its corresponding reporter strain. Uppercase nucleotides match sequences found in ARS1. Lowercase nucleotides form the spacers between adjacent domain A segments. The ARS consensus sequence is boxed and mutations are shown in bold.



**Fig. 3.** Secondary screen for hybrid constructs that bind the ARS consensus sequence. Plasmids recovered from six positive colonies identified in the initial screen of the hybrid expression library (Fig. 1A) were each introduced into two different reporter strains (Fig. 1B): JLY363, which contains a reporter gene with four copies of the wild-type ACS, and JLY365, which contains a reporter gene bearing four copies of a multiply mutated ACS (10). Each plasmid is represented by eight transformants within a rectangular box. The top row (+) of each box represents four independent transformants from JLY363 (wt ACS); the bottom row (-) represents four independent transformants from JLY365 (multiply mutated ACS).

hybrids were not responsible for the transcriptional induction attributed to these clones. Because of this finding and the lack of proper sequence specificity for the ACS element, AAP2-4 were not studied further.

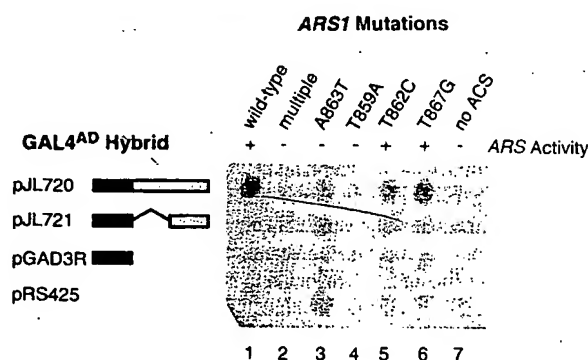
The full-length gene for AAP1 was cloned from a yeast genomic library and sequenced (15). AAP1 contains an open reading frame for a protein 435 amino acids long with a predicted molecular weight of 50,302 daltons (Fig. 5). The hybrid GAL4<sup>AD</sup>-AAP1 protein obtained from the screen was a fusion of the GAL4<sup>AD</sup> to the COOH-terminal two-thirds of the predicted full-length protein, indicating that this portion of the molecule is sufficient for association with the ACS.

When we were characterizing the AAP1 gene, the purification of a multi-protein complex that footprints the consensus sequence of several ARS elements, *in vitro* was reported (16). This complex, the origin recognition complex (ORC), contains six protein subunits with molecular weights of 50, 53, 56, 62, 72, and 120 kD. Comparison of peptide sequences from the 50-kD subunit of ORC (17) with the predicted protein sequence from AAP1 demonstrated that our gene encodes this subunit (Fig. 5) and confirmed the association between the AAP1 protein and the ACS. Because of this identity, we have renamed the AAP1 gene ORC6.

Although the specific interaction of the ORC6 protein (Orc6p) with the ACS suggests that it is part of a yeast initiator, the predicted Orc6p protein sequence reveals little about its function. Orc6p shows no significant similarity to any protein or translated open reading frame in the NCBI database. No matches to nucleotide binding (18) or helicase (19) motifs are present. The predicted amino acid sequence does not resolve whether Orc6p associates directly or indirectly with the ACS, because the protein contains no apparent DNA binding motifs.

On the other hand, the sequence of ORC6 hints at a possible connection with the regulatory machinery governing cell cycle progression. Orc6p contains four potential phosphorylation sites, (S/T)PXX, for cyclin-dependent protein kinases (20) clustered in the first half of the molecule (Fig. 5). Use of the more relaxed consensus site (S/T)P adds two more potential sites to this cluster. Although we have observed Orc6p phosphorylated *in vivo* on serine and threonine residues (13), we do not yet know the functional consequences of these modifications. However, because the initiation of yeast DNA replication commences promptly in response to the activation of this protein kinase in G1, it is tempting to speculate that Orc6p and possibly other ORC subunits are regulated substrates of

**Fig. 4.** The induction of *lacZ* by the AAP1 hybrid clone is dependent on a functional ACS and the AAP1 hybrid protein. Four hybrid protein constructs (rows) were transformed into a set of reporter strains (columns) and assayed for  $\beta$ -galactosidase activity (11). The reporter genes in these strains contain two direct repeats of the ACS sequences listed in Fig. 2. Columns 1 to 7 correspond to JLY360, JLY361, JLY429, JLY431, JLY433, JLY435, and pBgl-*lacZ*. ARS activity (12): + (50 to 100% of wild-type) or - (unmeasurable activity). The slightly higher level of *lacZ* activity associated with A863T is independent of the hybrid protein construct and probably represents background activity.



this kinase. Finally, as expected for a protein participating in nuclear events, Orc6p contains a potential nuclear localization signal (NLS) within the (S/T)PXX cluster and one within the COOH-terminal domain (Fig. 5). Orc6p can be detected in the nucleus when examined by immunofluorescence, but so far, only when the protein is significantly overexpressed (13).

Many yeast genes involved in DNA replication possess transcriptional control elements [MCB boxes (20a)] that direct their periodic expression in late G1. These elements are uniformly found within 250 nucleotides of the translational start codons for these genes. ORC6 exhibits one perfect and one near match to the MCB element 5' of its coding sequence, but the closest element is approximately 450 nucleotides from the predicted translational start for ORC6. Hence, though vaguely suggestive, the nucleotide sequence gives no strong indication that ORC6 belongs to the MCB class of genes.

We anticipated that if ORC6 was indeed involved in replication, it would be essential for viability. A marked deletion of the ORC6 gene (pJL731) (21) that removes all but 13 codons from the open reading frame was introduced into diploids from three dif-

ferent strain backgrounds. The resulting heterozygous ORC6 deletion strains, JLY481, JLY475, and JLY469, were induced to undergo meiosis, and 20 tetrads of each strain were dissected (21). In all backgrounds, the ORC6 disruption cosegregated with inviability, demonstrating that ORC6 is essential for cell growth. Microscopic examination revealed that mutant spores from JLY481 and JLY475 germinated, completed one to two rounds of cell division, and then arrested with a uniform large bud morphology reminiscent of cell division cycle mutants defective in DNA replication or nuclear division (22). The position of cell cycle arrest could not be established because the DNA content of these cells could not be readily measured. For unknown reasons, mutant spores derived from JLY469 germinated poorly.

The interpretation of these ORC6 deletion experiments was complicated by the presence of a second open reading frame (ORF2) of 250 amino acids on the antisense strand of the ORC6 gene. ORF2 spans nucleotides 1617 to 868 of the GenBank sequence and overlaps the COOH-terminal two-thirds of the ORC6 coding sequence. A marked deletion that removed the NH<sub>2</sub>-terminal third of the ORC6 coding sequence without affecting ORF2

**Fig. 5.** Predicted amino acid sequence of AAP1-ORC6. Shown in bold are the amino acid matches to peptide sequences from the ORC6 subunit (17). Matches to the consensus phosphorylation site (S/T)PXX of cyclin-dependent protein kinases (20) are underlined. The GAL4<sup>AD</sup>-ORC6 hybrid pJL720 contains amino acid residues 135 to 435; hybrid pJL721 contains residues 349 to 435 (14). Potential nuclear localization signals are at amino acid residues 117 to 122 and 263 to 279.

|             |            |            |            |            |
|-------------|------------|------------|------------|------------|
| 10          | 20         | 30         | 40         | 50         |
| MSMQVQHC    | AEVLRLDPQE | KPDWSSGYLK | KLNTATSILY | NTSLNKVMLK |
| 60          | 70         | 80         | 90         | 100        |
| QDEEVARCHI  | CAYIASQKMN | EKHMPDLCCY | IDSIPLEPKK | AKHLMNLFQ  |
| 110         | 120        | 130        | 140        | 150        |
| SLSNSSEPMKQ | FANTPSEKKN | KRSPVKNNGR | FTSSDPKELR | NQLFGTPTKV |
| 160         | 170        | 180        | 190        | 200        |
| RKSQNDSEFV  | IPELPPMQTN | ESPSITRRKL | AFEEDDEDE  | EEPGNDGLSL |
| 210         | 220        | 230        | 240        | 250        |
| KSHSNKSGIT  | TRNVDSEDE  | NHESDPTSE  | EPLGVQESRS | GRTKQNKAVG |
| 260         | 270        | 280        | 290        | 300        |
| KPQSELKTAK  | ALRKRGRIPN | SLLVKKYCKM | TTEIIIRLCN | DFELPREVAY |
| 310         | 320        | 330        | 340        | 350        |
| KIVDEYNINA  | SRLVCPQLV  | CGLVLNCTFI | VFNERRKDP  | RIDHFIVSKM |
| 360         | 370        | 380        | 390        | 400        |
| CSLMLTSKVD  | DVIECVKLVK | ELIIGEKWFR | DLQIRYDDFD | GIRYDEIIFR |
| 410         | 420        | 430        | 435        |            |
| KLGSMLQTTN  | ILVTDDQYNI | WKKRIEMDLA | LTEPL      |            |



**Table 1.** Viability of *cdc* mutants in the presence of high levels of *ORC6* expression. JL749 (GALp-HA-*ORC6*), JL772 (GALp-HA), and RS425 were introduced into each *cdc* mutant, and examined for growth at various temperatures under conditions that induce expression of *ORC6* (28, 29). Plus indicates mutants whose restrictive temperature remains unchanged in the presence of JL749 relative to JL772 and RS425. Minus indicates mutants whose restrictive temperature is lowered 5° to 7°C when JL749 is present.

| Strain | <i>cdc</i> Mutant | Viability |
|--------|-------------------|-----------|
| RDY488 | Wild-type         | +         |
| RDY501 | <i>cdc28-1</i>    | +         |
| RDY510 | <i>cdc4-1</i>     | +         |
| RDY664 | <i>cdc34-2</i>    | +         |
| RDY543 | <i>cdc7-4</i>     | +         |
| JLY310 | <i>cdc6-1</i>     | -         |
| JLY179 | <i>cdc46-1</i>    | -         |
| JLY338 | <i>cdc2-1</i>     | +         |
| JLY353 | <i>cdc17-1</i>    | +         |
| RDY619 | <i>cdc15-2</i>    | +         |

(pJL733) was introduced into diploids (21). Tetrad analysis again showed that the *ORC6* deletion cosegregated with cell death. Finally, an *ORC6* gene was constructed that contains a silent codon change for the *ORC6* ORF but introduces a UGA stop codon in ORF2 (Fig. 5) (23). This gene was able to rescue a haploid strain containing a full deletion of the *ORC6* ORF. We conclude that *ORC6* is essential for cell viability.

Our results validate the one-hybrid system screen as a method to identify and clone genes encoding proteins that recognize a DNA sequence of interest. An independent development of this screen has also been successful in identifying DNA-binding proteins (24), and a variation of this screen has been used to identify a binding site for a suspected DNA-binding protein (25). The one-hybrid approach is particularly useful for proteins that are difficult to detect biochemically or for which starting material for a purification is difficult to obtain.

Although *ORC* is a prime candidate for a yeast initiator protein because of its specific recognition of the ACS, biochemical analysis of *ORC* has yet to reveal additional properties that are characteristic of established initiator complexes from viral and bacterial systems (for example, adenosine triphosphate hydrolysis, DNA unwinding, and DNA helicase activity) (1). On the basis of our current biochemical understanding of *ORC*, one cannot rule out an alternative role for *ORC*, such as the repression of unscheduled initiation of DNA replication. Our isolation of the gene for *ORC6p*, as well as that reported for *ORC2p* (17, 26) should facilitate the functional analysis of *ORC*. We have demonstrated

that *ORC6* is essential for viability and is required at a specific stage of the cell cycle, as expected for a gene involved in the initiation of DNA replication. Further insight into the function of *ORC* will arise from studies on conditional *orc6* mutants, similar to those reported for *orc2* (17).

The identification of *ORC6* through the one-hybrid screen was based on the premise that the GAL4<sup>AD</sup>-*ORC6* hybrid can interact with the ARS consensus sequence in vivo. This interaction may reflect the incorporation of the hybrid into an origin recognition complex and the association of the entire complex with the ACS, because protein-DNA crosslinking experiments performed with *ORC* (16) suggest that *Orc6p* may not directly participate in DNA recognition. If so, the behavior of the *ORC6* hybrid in the screen provides the most compelling evidence that *ORC* binds to the origin of replication in vivo, a result first indicated by deoxyribonuclease protection studies on yeast origins in isolated nuclei (27).

Ultimately, definitive proof for a direct role of *ORC* in the initiation of DNA replication will require the establishment of an in vitro DNA replication system dependent on yeast origins, *ORC* protein, and quite possibly additional proteins. We have begun a search for such additional proteins by looking for genes that interact genetically with *ORC6*. Because germinating spores bearing an *ORC6* deletion appeared to exhibit a cell division cycle phenotype, we initially focused our attention on established *cdc* mutants. pJL749 (28), a plasmid that overexpresses *Orc6p* several hundred-fold (13), was introduced into a virtually isogenic set of temperature-sensitive *cdc* mutants arresting at various points in the cell cycle (29). Overexpression of *ORC6* selectively affected *cdc6* and *cdc46* mutants, lowering their restrictive temperature by 5° to 7°C; there was no significant effect on the other mutants examined or on the wild-type strain (Table 1).

Both *CDC6* and *CDC46* are required late in G1 for entry into S phase (30). Mutations in these genes interfere with the autonomous maintenance of ARS-containing plasmids, presumably by perturbing the replication of these plasmids (31). This plasmid-loss phenotype can be suppressed in *cdc6* by the addition of ARS elements to the plasmid (31). These results support the notion that *CDC6* and *CDC46* play a role in the proper function of origins early in replication. The specific interaction between these genes and *ORC6* suggests that *ORC* functions at a common step with *CDC6* and *CDC46*. It seems reasonable to speculate that *Cdc46p* or *Cdc6p* or both may help trigger the activity of *ORC* during the initiation of DNA replication and en-

sure that *ORC* acts at the proper time in the cell cycle.

## REFERENCES AND NOTES

1. T. J. Kelly, *J. Biol. Chem.* **263**, 17889 (1988); K. J. Mariani, *Annu. Rev. Biochem.* **61**, 673 (1992); A. Komberg, T. A. Baker, *DNA Replication* (Freeman, New York, 1992); B. Stillman, *Annu. Rev. Cell Biol.* **5**, 197 (1989).
2. M. L. DePamphilis, *Annu. Rev. Biochem.* **62**, 29 (1993).
3. J. L. Campbell and C. S. Newlon, in *The Molecular and Cellular Biology of the Yeast Saccharomyces*, J. R. Broach, J. R. Pringle, E. W. Jones, Eds. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1991), vol. 1, pp. 41-146.
4. W. L. Fangman and B. J. Brewer, *Annu. Rev. Cell Biol.* **7**, 375 (1991).
5. J. R. Broach et al., *Cold Spring Harbor Symp. Quant. Biol.* **47**, 1165 (1983); J. V. Van Houten and C. S. Newlon, *Mol. Cell. Biol.* **10**, 3917 (1990).
6. Y. Marahrens and B. Stillman, *Science* **255**, 817 (1992).
7. S. Fields and O.-K. Song, *Nature* **340**, 245 (1989); C.-T. Chien, P. T. Bartel, R. Stenglanz, S. Fields, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 9578 (1991).
8. R. Brent and M. Ptashne, *Cell* **43**, 729 (1985).
9. Three related hybrid expression libraries, YL1-3 (7), were a gift from S. Fields. The NH<sub>2</sub>-terminal portions of these hybrids consist of the SV40 nuclear localization signal and amino acids 768 to 881 of the GAL4 activation domain (GAL4<sup>AD</sup>). The COOH-terminal portions were derived from random yeast protein segments which have been fused to the end of the GAL4<sup>AD</sup>. These segments are encoded by short (1-3 kb) fragments from a Sau 3A partial digest of yeast genomic DNA. Together, YL1-3 ensure that all three reading frames of these fragments can be expressed.
10. pLR1Δ1 [R. W. West Jr., R. R. Rogers, M. Ptashne, *Mol. Cell. Biol.* **4**, 2467 (1984)] was a gift of R. West. We generated pBgl-lacZ from pLR1Δ1 by: (i) substituting an Xho I-Bgl II-Xho I polylinker 5'-CCTCGAGGAGATCTCTCGAGG-3' for the Xho I linker and (ii) precisely excising a Hind III fragment containing 2-μm sequences. The resulting vector has a unique Bgl II site approximately 100-bp upstream of the TATA box for insertion of DNA sequences in the promoter region and a unique Stu I site for targeted integration of the plasmid at the URA3 locus. Multiple direct repeats of ARS1 domain A and several of its mutant derivatives were inserted into the Bgl II site of pBgl-lacZ to generate all the reporter genes used in this work. The inserted repeat elements, derived from complementary oligonucleotides shown in Fig. 2, were oriented with the TATA box to their right. Each reporter gene construct was integrated into the URA3 locus of GGY1 (*MATα* *Agal4* *Agal80* *ura3* *leu2* *his3* *ade2* *tyr*) [G. Gill and M. Ptashne, *Cell* **51**, 121 (1987)] to create a reporter strain (Fig. 4). Integration of pBgl-lacZ into GGY1 generated JLY387.
11. YEPD (rich complete) and SD (synthetic dropout) media are as described [J. B. Hicks and I. Herskowitz, *Genetics* **83**, 245 (1976)]. Standard methods were used for manipulation of yeast cells [C. Guthrie and G. R. Fink, Eds., *Guide to Yeast Genetics and Molecular Biology* (Academic Press, San Diego, 1991)] and DNA [F. M. Ausubel et al., Eds., *Current Protocols in Molecular Biology* (Wiley, New York, 1989)]. Libraries YL1-3 were transformed [R. H. Schiestl and R. D. Gaitz, *Current Genetics* **16**, 339 (1989)] into JLY363 (10) and plated on SD-Leu at a density of 2 to 5000 colonies per 10-cm plate. Five hundred thousand transformants were obtained for YL1 and YL2, and 200,000 for YL3. Transformants were assayed on filters for production of β-galactosidase [L. Breeden and K. Nasmyth, *Cold Spring Harbor Symp. Quant. Biol.* **47**, 643 (1985)]. Forty-nine isolates remained positive after colony purification (15 from YL-1, 22 from YL-2, and 12 from YL-3), and library plasmids were extracted from them.

- These plasmids were each transformed into both JLY363 and its mutant counterpart JLY365 (10). Nine plasmids induced greater  $\beta$ -galactosidase activity in the wild-type reporter strain than the control. These plasmids were classified into five clones, AAP1 through AAP5, on the basis of their Hind III restriction pattern. Each clone was retested in JLY360, JLY361, JLY387, JLY429, JLY431, JLY433, JLY435 (Fig. 4). The AAP1 hybrid clone was called pJL720. The AAP1 gene was later renamed *ORC6*.
12. The *ARS* function of the sequences in Fig. 4 was analyzed in the context of *ARS1* domain B (Bgl II-Hinf I fragment; nt 853-734) in the following CEN-based URA3-containing plasmids: pJL347 (wt), pJL243 (multiple), pJL326 (A863T), pJL338 (T869A), pJL330 (T862C), and pJL316 (T867G). These plasmids were transformed into JLY106 (*MAT $\alpha$  ura3 leu2 his3 trp1 lys2 ade2*) and its homozygous diploid counterpart JLY162. pJL243, pJL326, and pJL338 did not yield a high frequency of transformation and could not be assayed quantitatively for *ARS* function. pJL347, pJL330, and pJL316 transformed cells with high efficiency and were assayed for mitotic stability [D. T. Stinchcomb, K. Struhl, R. W. Davis, *Nature* 282, 39 (1979)].
  13. J. J. Li and I. Herskowitz, unpublished material.
  14. The *ORC6* hybrid construct originally isolated from the YL3 library (pJL720) has two Bam HI sites. The 5' site, which is created by the hybrid junction, corresponds to the Sau 3A site at nucleotide 843. Excision of the segment between the two sites generated pJL721, leaving amino acid residues 339 to 435 in frame with the *GAL4<sup>AD</sup>* (Fig. 5). pGAD3R (11), the parent vector for the YL3 library, contains no *ORC6* sequence. pRS425 [T. W. Christianson, R. S. Sikorski, M. Dante, J. H. Shero, P. Hieter, *Gene* 110, 119 (1992)] contains no components of the fusion protein.
  15. All sequencing was performed with Sequenase (USB) on collapsed double-stranded templates. The protein coding segments of the AAP1 through AAP5 hybrid clones were sequenced from their junction with the *GAL4<sup>AD</sup>* to their stop codon. Two of the *ORC6* sequencing primers were used as colony hybridization probes to screen a high copy number yeast genomic library [M. Carlson and D. Botstein, *Cell* 28, 145 (1982)] for a clone of the full-length *ORC6* gene (pJL724). The full-length gene was sequenced on both strands with oligonucleotide primers positioned approximately 200 nucleotides apart. The accession number for the *ORC6* sequence reported in this paper is L23323.
  16. S. P. Bell and B. Stillman, *Nature* 357, 128 (1992).
  17. S. P. Bell, R. Kobayashi, B. Stillman, *Science* 262, 1844 (1993).
  18. T. C. Hodgman, *Nature* 333, 22 (1988); J. E. Walker, M. Sarasté, M. J. Runswick, N. J. Gay, *EMBO J.* 1, 945 (1982).
  19. P. Linder et al., *Nature* 337, 121 (1989).
  20. E. A. Nigg, *Seminars in Cell Biology* 2, 261 (1991).
  - 20A. B. J. Andrews and S. W. Mason, *Science* 261, 1543 (1993).
  21. Marked *ORC6* deletions were constructed by replacing nucleotides 458-1721 (pJL731) or nucleotides 458-846 (pJL733) of the GenBank sequence with the URA3 Hind III fragment oriented in the opposite direction to that of the *ORC6* sequence. Each construct was used to generate heterozygous deletions of *ORC6* in diploid strains by one-step gene replacement. *ORC6* deletion analysis was performed in JLY461 (*MAT $\alpha$ /MAT $\alpha$  ura3/ura3 leu2/his3 trp1/trp1 ade2/ade2 [cir<sup>+</sup>]*), JLY462 (*MAT $\alpha$ /MAT $\alpha$  ura3/ura3 leu2/his3 trp1/trp1 his4/his4 can1/can1*), and JLY463 (*MAT $\alpha$ /MAT $\alpha$  ura3/ura3 leu2/his3 trp1/trp1 his3/his3*); their respective genetic backgrounds are S288c, EG123, and A364a. Disruption of JLY461, JLY462, and JLY463 by pJL731 (full deletion) created JLY481, JLY475, and JLY469, respectively. Disruption of JLY461, JLY462, and JLY463 by pJL733 (NH<sub>2</sub>-terminal deletion) created JLY485, JLY479, JLY473, respectively. These heterozygous marked deletion strains were sporulated, and 20 tetrads of each were dissected and grown on YEPD to assess viability.
  22. J. R. Pringle and L. H. Hartwell, in *The Molecular Biology of the Yeast Saccharomyces*, J. N. Strathern, E. W. Jones, J. R. Broach, Eds. (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1981), vol. 1, pp. 97-142.
  23. A point mutant (pJL766) was made by replacing the Bam HI-Sph I fragment of the full-length clone with a Bam HI-Sph I fragment generated by PCR from pJL720 with the primers  
5'-CAAGGATCCAGAATTGATCATTTATAGT-CAG-3'  
5'-GTTATAGGGCTAAAGGCATGC-3'.  
The mutation, shown in bold, changes nucleotide 1471 of the GenBank sequence from C to T and was confirmed by sequence analysis.
  24. M. M. Wang and R. R. Reed, *Nature* 364, 121 (1993).
  25. T. E. Wilson, T. J. Fahrner, M. Johnston, J. Milbrant, *Science* 252, 1296 (1991).
  26. M. Foss, F. J. McNally, P. Laurensen, J. Rine, *ibid.* 262, 1838 (1993).
  27. J. F. X. Diffley and J. H. Cocker, *Nature* 357, 169 (1992).
  28. pJL749 contains the *GAL1* promoter (nucleotides 146 to 816) driving the expression of *ORC6* (nucleotides 443 to 2298) in the high-copy yeast shuttle vector RS425 [T. W. Christianson, R. S. Sikorski, M. Dante, J. H. Shero, P. Hieter, *Gene* 110, 119 (1992)]. The sequence 5'-CCCCGATCCC ATG GCC TAC CCA TAT GAT GTT CCA GAT TAC GCT TCT TTG GGT CCA GGG CTG CAG GAA TTC GGG CCC ATC-3' lies between the *GAL1* promoter and *ORC6* and contains the influenza hemagglutinin (HA) epitope fused to the NH<sub>2</sub>-terminus of *ORC6*. This construct complements a deletion of the *ORC6* gene. pJL772 is identical to pJL749 except that it lacks the *ORC6* sequence.
  29. The *cdc* mutant strains listed in Table 1 have been backcrossed four to five times against two con-
- genic strains derived from A364a, *RDY487* (*MAT $\alpha$  leu2 ura3 trp1*) and *RDY488* (*MAT $\alpha$  leu2 ura3 trp1*). All are *ura3 leu2 trp1*. *RDY510*, *RDY664*, *JLY310*, and *JLY179* are *MAT $\alpha$* ; the rest are *MAT $\alpha$* . Additional markers can be found in JLY310 (*ade2*), *RDY543* (*his3*), and *RDY619* (*pep4 $\Delta$ ::TRP1 his3 ade2*). The *RDY* strains were a gift from R. Deshaies. pJL749, pJL772, and RS425 (28) were transformed into these strains and plated on SD-Leu at 22°C. Four colony-purified isolates from each transformation were patched onto SD-Leu plates and replica-plated to SGAL-Leu plates, all at 22°C. The patches on SGAL-Leu were replicated to a series of prewarmed SGAL-Leu plates at 22°, 25°, 27°, 30°, 32.5°, 35°, 37°, and 38°C. The viability of *cdc* mutants containing pJL749 was compared to those containing pJL772 and pRS425.
30. L. H. Hartwell, *J. Mol. Biol.* 104, 803 (1976); K. M. Hennessy, C. D. Clark, D. Botstein, *Genes Dev.* 4, 2252 (1990).
  31. Y. Chen, K. M. Hennessy, D. Botstein, B.-K. Tye, *Proc. Natl. Acad. Sci. U.S.A.* 89, 10459 (1992); E. Hogan and D. Koshland, *ibid.* 89, 3098 (1992).
  32. We thank C. Peterson for preparation of library DNA, F. Banuett for oligonucleotide synthesis, and A. Lynn for superb help with the figures. We appreciate S. Fields, N. Hollingsworth, A. Sil, R. Deshaies, P. Sorger, P. Jackson, S. Sanders, A. Johnson, C. Detweiler, and A. Lynn for helpful discussions or useful suggestions on the manuscript. This work was supported by NIH grant AI18738 (to I.H.). J.J.L. is a Lucille P. Markey Scholar and this work was supported in part by a grant from the Lucille P. Markey Charitable Trust. J.J.L. also appreciates the early support from Bristol-Myers Squibb through the Life Sciences Research Foundation. The accession number for the *ORC6* sequence reported in this paper is L23323.

13 October 1993; accepted 17 November 1993

## Sharing of the Interleukin-2 (IL-2) Receptor $\gamma$ Chain Between Receptors for IL-2 and IL-4

Motonari Kondo, Toshikazu Takeshita, Naoto Ishii, Masataka Nakamura, Sumiko Watanabe, Ken-ichi Arai, Kazuo Sugamura\*

The  $\gamma$  chain of the interleukin-2 (IL-2) receptor is an indispensable subunit for IL-2 binding and intracellular signal transduction. A monoclonal antibody to the  $\gamma$  chain, TUGm2, inhibited IL-2 binding to the functional IL-2 receptors and also inhibited IL-4-induced cell growth and the high-affinity binding of IL-4 to the CTLL-2 mouse T cell line. Another monoclonal antibody, TUGm3, which reacted with the  $\gamma$  chain cross-linked with IL-2, also immunoprecipitated the  $\gamma$  chain when cross-linked with IL-4. These results suggest that the IL-2 receptor  $\gamma$  chain is functionally involved in the IL-4 receptor complex.

**Functional high-affinity receptors for cytokines are generally complexes consisting of binding subunits ( $\alpha$  chains) with low affinities to ligands and effector subunits ( $\beta$  chains) to transduce signals, both of which are members of the cytokine receptor super-**

**family (1). The same  $\beta$  chain is shared by receptors for IL-3, IL-5, and granulocyte-macrophage colony-stimulating factor (GM-CSF) (1, 2). Another molecule, gp130, is shared as a signaling molecule by the receptors for IL-6, leukemia inhibitory factor (LIF), oncostatin M (OSM), and ciliary neurotrophic factor (CNTF) receptors (3). The IL-2 receptor is also a complex (4), but the low-affinity  $\alpha$  chain is not a member of the cytokine receptor superfamily and the high-affinity receptor contains, in addition to the  $\alpha$  and  $\beta$  chains, the  $\gamma$**

M. Kondo, T. Takeshita, N. Ishii, M. Nakamura, K. Sugamura, Department of Microbiology, Tohoku University School of Medicine, Sendai 980, Japan. S. Watanabe and K.-i. Arai, Department of Molecular and Developmental Biology, Institute of Medical Science, University of Tokyo, Tokyo 108, Japan.

\*To whom correspondence should be addressed.



# Components of the Arabidopsis C-Repeat/Dehydration-Responsive Element Binding Factor Cold-Response Pathway Are Conserved in *Brassica napus* and Other Plant Species<sup>1</sup>

Kirsten R. Jaglo<sup>2</sup>, Susanne Kleff<sup>3</sup>, Keenan L. Amundsen, Xin Zhang<sup>4</sup>, Volker Haake, James Z. Zhang, Thomas Deits, and Michael F. Thomashow\*

Department of Crop and Soil Science, Michigan State University, East Lansing, Michigan 48824 (K.R.J., K.L.A., X.Z., M.F.T.); MBI International, Lansing, Michigan 48909 (S.K., T.D.); Mendel Biotechnology Inc., Hayward, California 94545 (V.H., J.Z.Z.); and Michigan State University-Department of Energy Plant Research Laboratory (M.F.T), Michigan State University, East Lansing, Michigan 48824

Many plants increase in freezing tolerance in response to low, nonfreezing temperatures, a phenomenon known as cold acclimation. Cold acclimation in *Arabidopsis* involves rapid cold-induced expression of the C-repeat/dehydration-responsive element binding factor (CBF) transcriptional activators followed by expression of CBF-targeted genes that increase freezing tolerance. Here, we present evidence for a CBF cold-response pathway in *Brassica napus*. We show that *B. napus* encodes CBF-like genes and that transcripts for these genes accumulate rapidly in response to low temperature followed closely by expression of the cold-regulated *Bn115* gene, an ortholog of the *Arabidopsis* CBF-targeted *COR15a* gene. Moreover, we show that constitutive overexpression of the *Arabidopsis* CBF genes in transgenic *B. napus* plants induces expression of orthologs of *Arabidopsis* CBF-targeted genes and increases the freezing tolerance of both nonacclimated and cold-acclimated plants. Transcripts encoding CBF-like proteins were also found to accumulate rapidly in response to low temperature in wheat (*Triticum aestivum* L. cv Norstar) and rye (*Secale cereale* L. cv Puma), which cold acclimate, as well as in tomato (*Lycopersicon esculentum* var. Bonny Best, Castle Mart, Micro-Tom, and D Huang), a freezing-sensitive plant that does not cold acclimate. An alignment of the CBF proteins from *Arabidopsis*, *B. napus*, wheat, rye, and tomato revealed the presence of conserved amino acid sequences, PKK/RPAGR<sub>x</sub>KFxETRHP and DSAWR, that bracket the AP2/EREBP DNA binding domains of the proteins and distinguish them from other members of the AP2/EREBP protein family. We conclude that components of the CBF cold-response pathway are highly conserved in flowering plants and not limited to those that cold acclimate.

Plants vary greatly in their abilities to survive freezing temperatures (Sakai and Larcher, 1987). Whereas plants from tropical regions have essentially no capacity to withstand freezing, herbaceous plants from temperate regions can survive freezing at temperatures ranging from  $-5$  to  $-30^{\circ}\text{C}$ , depending on

the species. It is significant that the maximum freezing tolerance of plants is not constitutive, but is induced in response to low temperatures (below approximately  $10^{\circ}\text{C}$ ), a phenomenon known as "cold acclimation" (Hughes and Dunn, 1996; Thomashow, 1999). Nonacclimated wheat (*Triticum aestivum* L. cv Norstar) plants, for instance, are killed at freezing temperatures of about  $-5^{\circ}\text{C}$ , but after cold acclimation, can survive temperatures down to about  $-20^{\circ}\text{C}$ . Determining what accounts for the differences in freezing tolerance between plant species and the molecular basis of cold acclimation is of basic scientific interest and has the potential to provide new approaches to improve the freezing tolerance of plants, an important agronomic trait.

A recent advance in understanding cold acclimation in *Arabidopsis* was the discovery of the C-repeat/dehydration-responsive element binding factor (CBF) cold-response pathway (see Thomashow, 2001). *Arabidopsis* encodes a small family of cold-responsive transcriptional activators known either as CBF1, CBF2, and CBF3 (Stockinger et al., 1997; Gilmour et al., 1998) or DREB1b, DREB1c, and DREB1a (Liu et al., 1998; Kasuga et al., 1999), respec-

<sup>1</sup> This research was supported by a subcontract (no. 593-0219-06) under the U.S. Department of Agriculture/Cooperative State Research, Education, and Extension Service Cooperative Agreement North Central Biotechnology Initiative (no. 96-34340-2711), by Mendel Biotechnology, Inc., and by the Michigan Agricultural Experiment Station.

<sup>2</sup> Present address: Campus Box 0448, Department of Biochemistry and Biophysics, University of California, San Francisco, CA 94143-0448.

<sup>3</sup> Present address: 341 Food Safety Building, Michigan State University, East Lansing, MI 48824.

<sup>4</sup> Home institution: Horticultural Research Institute, Heilongjiang Academy of Agricultural Sciences, 368 Xuefu Road, Harbin 150086, China.

\* Corresponding author; e-mail thomash6@msu.edu; fax 517-353-9168.

Article, publication date, and citation information can be found at [www.plantphysiol.org/cgi/doi/10.1104/pp.010548](http://www.plantphysiol.org/cgi/doi/10.1104/pp.010548).

tively. The CBF transcription factors, which are members of the AP2/EREBP family of DNA-binding proteins (Riechmann and Meyerowitz, 1998), recognize the cold- and dehydration-responsive DNA regulatory element designated the CRT (C-repeat; Baker et al., 1994)/DRE (dehydration-responsive element; Yamaguchi-Shinozaki and Shinozaki, 1994). CRT/DRE elements, which have a conserved 5-bp core sequence of CCGAC, are present in the promoter regions of many cold- and dehydration-responsive genes of Arabidopsis including those designated COR (cold-regulated; Thomashow, 1999). The CBF genes are induced within 15 min of plants being exposed to low nonfreezing temperatures followed at about 2 h by induction of cold-regulated genes that contain the CRT/DRE-regulatory element, i.e. the "CBF regulon" (Gilmour et al., 1998; Liu et al., 1998). Over the next few days at low temperature, the plants increase in freezing tolerance reaching a maximum level within 1 to 2 weeks.

A role for the CBF regulon in the enhancement of freezing tolerance is indicated by the results of CBF overexpression experiments. Constitutive expression of the CBF genes in transgenic Arabidopsis plants results in the induction of COR gene expression and an increase in freezing tolerance without a low temperature stimulus (Jaglo-Ottosen et al., 1998; Liu et al., 1998; Kasuga et al., 1999; Gilmour et al., 2000). It is significant that multiple biochemical changes that are associated with cold acclimation and thought to contribute to increased freezing tolerance, including the accumulation of sugars and Pro, occur in nonacclimated transgenic Arabidopsis plants that constitutively express CBF3 (Gilmour et al., 2000). Thus, it has been proposed that the CBF genes act to integrate the activation of multiple components of the cold acclimation response (Gilmour et al., 2000).

The discovery of the Arabidopsis CBF cold-response pathway raises a number of fundamental questions about plant freezing tolerance. Do plants other than Arabidopsis have CBF genes that are cold induced? If so, do they activate expression of CBF regulons that increase freezing tolerance? Are cold-regulated orthologs of CBF genes limited to plants that cold acclimate? The results presented here begin to address these questions.

## RESULTS

### A CBF Cold-Response Pathway in *Brassica napus*

*B. napus*, like Arabidopsis, cold acclimates and is a member of the Cruciferae family. As a first step to determine whether *B. napus* has a cold-response pathway related to the CBF cold-response pathway of Arabidopsis, we asked whether *B. napus* encoded CBF-like proteins. The results indicated that it did. cDNA clones encoding two different CBF-like proteins (accession nos. AF370733 and AF370734) were identified by screening *B. napus* cDNA libraries using PCR-

generated probes (see "Materials and Methods"). The *B. napus* CBF-like proteins were 92% identical in amino acid sequence to each other and approximately 76% identical in sequence to Arabidopsis CBF1. An alignment of the *B. napus* proteins with Arabidopsis CBF1 indicated that the sequence identity extended throughout the protein, but was greatest in the AP2/EREBP DNA-binding domain (Fig. 1 includes an alignment of one *B. napus* CBF protein against Arabidopsis CBF1). A sequence for a third *B. napus* CBF polypeptide has been deposited by others (accession no. AF084185; N. Zhou, G. Wu, Y.-P. Gao, R.W. Wilen, and L.V. Gusta).

Transcripts encoding *B. napus* CBF-like proteins were found to accumulate rapidly (within 30 min) upon exposure of plants to low temperature (Fig. 2). This was closely followed by expression of *Bn115* (Weretilnyk et al., 1993), a cold-regulated ortholog of Arabidopsis *COR15a* (Hajela et al., 1990). Arabidopsis *COR15a* is cold regulated, has CRT/DRE regula-

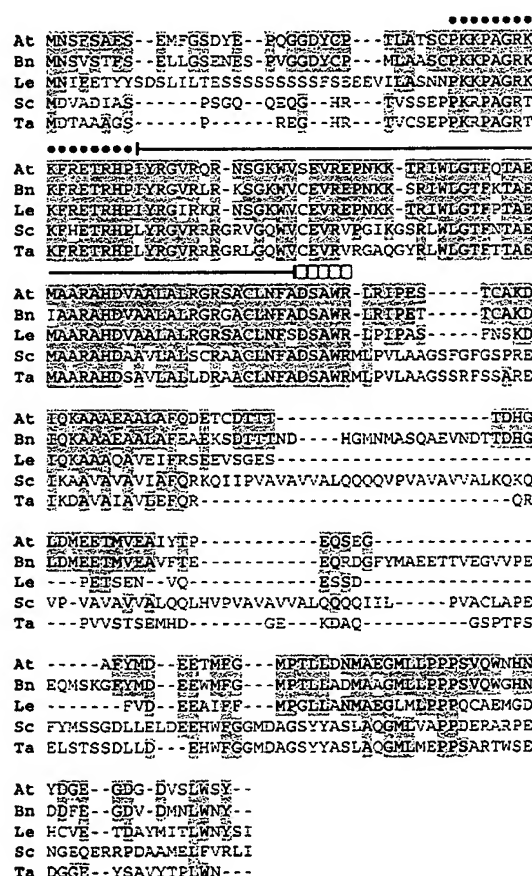
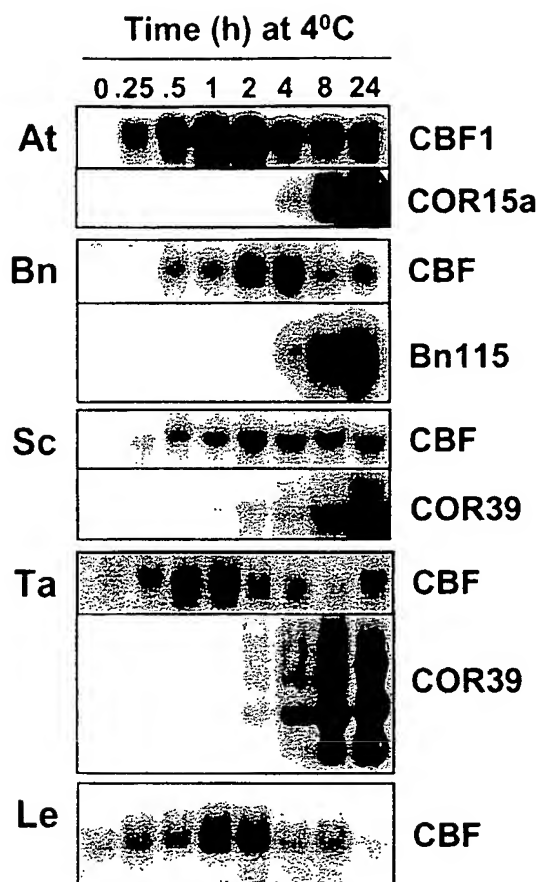


Figure 1. Alignment of CBF-like proteins. The amino acid sequence shown are for: At, Arabidopsis CBF1 (accession no. AAC49662); Bn, *B. napus* CBF (accession no. AF370733); Le, tomato (*Lycopersicon esculentum*) CBF (accession no. AY034473); Sc, rye (*Secale cereale*) CBF (accession no. AF370730); and Ta, wheat CBF (accession no. AF376136). The AP2/EREBP domain is indicated by an overline and the signature sequences PKK/RPAGRxFxETRHP and DSAWR are indicated by black circles and white boxes, respectively.



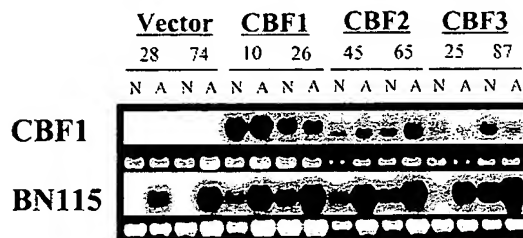
**Figure 2.** Accumulation of *CBF* and putative target gene transcripts in response to low temperature. Plants were grown at normal growth temperatures (20°C–22°C) and transferred to low temperature (4°C) for the indicated times. Total RNA was isolated from leaves and northern analyses performed using probes for *CBF* transcripts and putative *CBF*-targeted cold-regulated genes for *B. napus* (*Bn115*), wheat and rye (*Wcs120/COR39*), and Arabidopsis (*COR15a*) as described in "Materials and Methods." At, Arabidopsis; Bn, *B. napus*; Sc, rye; Ta, wheat; Le, tomato.

tory elements, and is induced in response to the *CBF* transcriptional activators (Gilmour et al., 1998; Jaglo-Ottosen et al., 1998). Cold-regulated expression of the *B. napus* *Bn115* gene involves a DNA regulatory element, the low temperature responsive element, that contains the CRT/DRE core sequence CCGAC (Jiang et al., 1996). As with Arabidopsis *CBF* transcripts, *B. napus* *CBF* transcripts reached maximum levels within a few hours of plants being transferred to low temperature, after which time they decreased, but at 24 h remained elevated over the level found in non-acclimated plants.

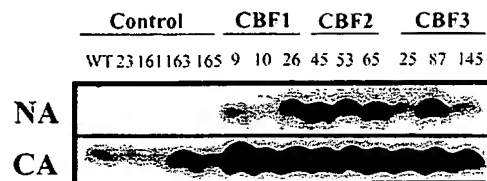
Constitutive expression of Arabidopsis *CBF1*, *CBF2*, or *CBF3* in transgenic Arabidopsis plants activates expression of the target CRT/DRE-containing *COR* genes and increases freezing tolerance without a low temperature stimulus (Gilmour et al., 1998; Jaglo-Ottosen et al., 1998; Liu et al., 1998; S.J. Gilmour and M.F. Thomashow, unpublished data). We rea-

soned that if *B. napus* had a similar *CBF*-like cold-response pathway, then expression of the Arabidopsis *CBF* genes in transgenic *B. napus* might also activate expression of *Bn115* and other cold-regulated genes containing the CRT/DRE-related regulatory elements and increase plant freezing tolerance. This was found to be the case. Constitutive expression of Arabidopsis *CBF1*, *CBF2*, and *CBF3* in transgenic *B. napus* caused the accumulation of transcripts for *Bn115* (Fig. 3A) and *Bn28* (not shown) without a low temperature stimulus; *Bn28* encodes an ortholog of the CRT/DRE-regulated cold-responsive gene *COR6.6* (Hajela et al., 1990). Immunoblot analysis further indicated that the *BN28* protein accumulated in nonacclimated plants that expressed *CBF1*, *CBF2*, or *CBF3* (Fig. 3B). Finally, the levels of the *BN28* protein were higher in cold-

#### A. *CBF* and *Bn115* transcript levels

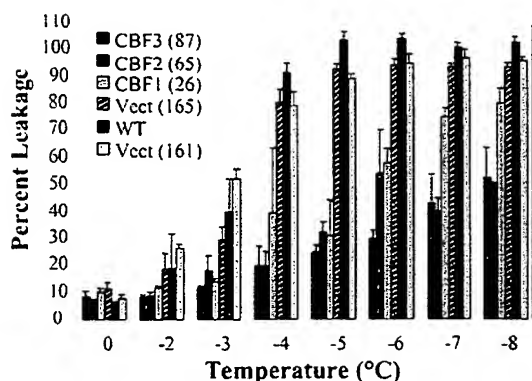


#### B. *BN28* protein levels



**Figure 3.** Effect of overexpressing Arabidopsis *CBF* genes in transgenic *B. napus* plants on expression of endogenous cold-regulated genes *Bn115* and *Bn28*. A, Transcript levels of the Arabidopsis *CBF* transgenes and the endogenous *B. napus* *Bn115* gene in control (vector) and *CBF*-expressing (*CBF1*, *CBF2*, and *CBF3*) *B. napus* transgenic plants that were either nonacclimated (N) or cold acclimated (A) for 3 weeks. Total RNA was isolated from pooled plants of the indicated transgenic lines and subjected to northern analysis using probes prepared from cDNAs for either the Arabidopsis *CBF1* gene or *B. napus* *Bn115* gene. Numbers above the samples refer to the specific transgenic lines tested. Loading controls show the 18S ribosomal RNA band from the corresponding ethidium bromide-stained agarose gel used for the northern analysis. B, Levels of the *B. napus* *BN28* protein in nonacclimated (NA) and cold-acclimated (CA) control and *CBF*-expressing transgenic *B. napus* plants. Total soluble protein (100  $\mu$ g) prepared from nonacclimated and 3-week cold-acclimated plants was subjected to immunoblot analysis using antiserum raised to the *BN28* polypeptide (Boothe et al., 1997). Numbers above each sample refer to the specific transgenic line tested. The sample designated WT was from plants that had not been transformed. Protein transfer for line 10 was inefficient in this experiment due to a bubble in the gel.

## A. Nonacclimated Plants



## B. Cold-Acclimated Plants

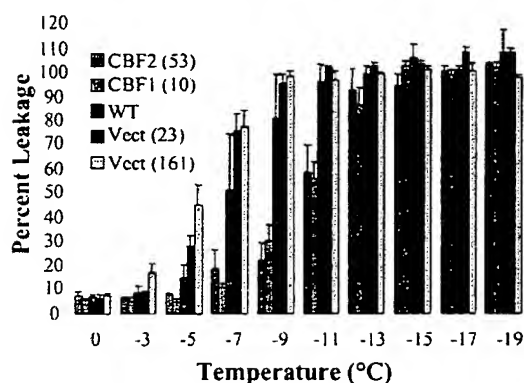


Figure 4. Freezing tolerance of leaf tissue from nonacclimated (A) or cold-acclimated (B) control and CBF-expressing *B. napus* plants. Leaves from nonacclimated and cold-acclimated seedlings were frozen to the temperatures indicated and cellular damage assessed by measuring electrolyte leakage as described in "Materials and Methods." Numbers in parentheses indicate the specific transgenic lines tested. Error bars indicate the SDs of the three replicates of each data point.

acclimated CBF-expressing plants than they were in control plants (Fig. 3B).

Electrolyte leakage experiments indicated that expression of the Arabidopsis CBF genes in *B. napus* resulted in an increase in freezing tolerance. In the experiment shown in Figure 4A, leaf tissue from nonacclimated control *B. napus* plants had  $EL_{50}$  values (the freezing temperature that causes leakage of 50% of total electrolytes) between  $-3^{\circ}\text{C}$  and  $-4^{\circ}\text{C}$ , whereas the leaf tissue of plants expressing CBF1, CBF2, or CBF3 had  $EL_{50}$  values of about  $-6^{\circ}\text{C}$ . Combined results from multiple electrolyte leakage experiments indicated that leaf tissue from nonacclimated control *B. napus* plants had an  $EL_{50}$  value of about  $-2.1^{\circ}\text{C}$ , whereas leaf tissue from nonacclimated CBF-expressing plants had an  $EL_{50}$  value of about  $-4.7^{\circ}\text{C}$  (Table I). CBF expression was also found to cause an increase in the freezing tolerance of cold-acclimated plants. In the experiment shown in Figure 4B, leaf

tissue from cold-acclimated control *B. napus* plants had  $EL_{50}$  values of about  $-6^{\circ}\text{C}$ , whereas the leaf tissue of plants expressing either CBF1 or CBF2 had  $EL_{50}$  values of about  $-11^{\circ}\text{C}$ . Combined results from multiple experiments indicated that leaf tissue from cold-acclimated control *B. napus* plants had an  $EL_{50}$  value of about  $-8.1^{\circ}\text{C}$ , whereas leaf tissue from cold-acclimated CBF-expressing plants had an  $EL_{50}$  value of  $-12.7^{\circ}\text{C}$  (Table I).

## Cold-Responsive CBF-Like Genes in Wheat and Rye

The results presented above indicated that *B. napus* encodes a CBF cold-response pathway related to that found in Arabidopsis. We next asked whether more distantly related plants that cold acclimate have CBF-like genes that are rapidly induced in response to low temperature. cDNA libraries of rye and wheat were screened for clones encoding CBF-like proteins using probes generated by PCR (see "Materials and Methods"). This resulted in the identification of cDNA inserts encoding one wheat (accession no. AF376136) and three rye (accession nos. AF370728, AF370729, and AF370730) CBF-like polypeptides. The rye and wheat polypeptides shared 30% to 34% sequence identity with Arabidopsis CBF1, most of which was due to a high degree of identity between the AP2/EREBP DNA-binding domains (Fig. 1 includes an alignment of the wheat and a rye CBF protein with Arabidopsis CBF1). However, a striking feature of the wheat and rye proteins was that they had in common with the Arabidopsis and *B. napus* CBF proteins short polypeptide sequences that flanked the AP2/EREBP sequence; PKK/RPAGR $\times$ KF $\times$ ETRHP immediately upstream of the AP2/EREBP domain and the sequence DSAWR just downstream from it (see Fig. 1). It is significant that of the more than 140 AP2/EREBP domain proteins predicted to be encoded by Arabidopsis (Rie-

Table I. Freezing tolerance ( $EL_{50}$  values in  $^{\circ}\text{C}$ ) for nonacclimated and cold-acclimated control and CBF-expressing transgenic *B. napus* plants<sup>a</sup>

| Plants         | Nonacclimated        | Cold Acclimated       |
|----------------|----------------------|-----------------------|
| Control        | $-2.1 \pm 0.34$ (10) | $-8.1 \pm 0.42$ (8)   |
| CBF expressing | $-4.7 \pm 0.40$ (23) | $-12.7 \pm 0.52$ (12) |

<sup>a</sup>  $EL_{50}$  values were calculated using combined data from individual nonacclimated or cold-acclimated control and CBF-expressing plants (no. of plants used are indicated in parentheses). All values were significantly different from each other ( $P < 0.001$ ) as determined by ANOVA. Nonacclimated control plants used were: wild type (2), vector-23 (2), vector-161 (4), vector-163 (1), and vector-165 (1). Nonacclimated CBF-expressing plants used were: CBF1-9 (1), CBF1-10 (3), CBF1-26 (3), CBF2-45 (1), CBF2-53 (2), CBF2-65 (3), CBF3-25 (2), CBF3-87 (2), CBF3-108 (1), CBF3-129 (1), and CBF3-145 (3). Cold-acclimated control plants used were: wild type (2), vector-23 (1), vector-161 (3), vector-163 (1), and vector-165 (1). Cold-acclimated CBF-expressing plants used were: CBF1-9 (1), CBF1-10 (2), CBF1-26 (2), CBF2-45 (1); CBF2-53 (1); CBF2-65 (1); CBF3-25 (1); CBF3-87 (1); CBF3-145 (2).

hmann et al., 2000), only CBF1, CBF2, and CBF3 were found to have the PKK/RPAGR<sub>x</sub>KFxETRHP and DSAWR "signature sequences" surrounding the AP2/EREBP domain. The AP2/EREBP domains of three additional Arabidopsis AP2/EREBP proteins (accession nos. 3241926, AC025417, and AC010795) were also found bracketed by the nearly identical sequences PKK/RRAGR<sub>x</sub>FxETRHP and DSAWR.

As in Arabidopsis and *B. napus*, CBF-like transcripts accumulated rapidly (within 15–30 min) in response to low temperature in both wheat and rye (Fig. 2). This was followed at about 2 h by accumulation of transcripts for the cold-responsive *Wcs120/COR39* gene family (Guo et al., 1992; Houde et al., 1992; Fig. 2). *Wcs120/COR39*, which is an ortholog of the CBF-targeted cold-regulated *COR47* gene of Arabidopsis (Gilmour et al., 1992), is a potential CBF target because its promoter is activated in response to low temperature and has multiple copies of the CRT/DRE core sequence CCGAC (Ouellet et al., 1998).

#### Cold-Responsive CBF-Like Genes in Tomato

The results presented above supported the hypothesis that a common feature of cold acclimation is rapid cold induction of genes encoding CBF-like transcriptional activators. A fundamental question raised was whether plants that do not cold acclimate encode CBF-like proteins and whether transcripts encoding them accumulate rapidly in response to low temperature. A search of the public databases indicated that tomato encoded multiple AP2/EREBP proteins that share significant sequence identity with Arabidopsis CBF1. A clone for one expressed sequence tag (EST; accession no. AI89824) was obtained and the complete DNA sequence of the insert was determined (accession no. AY034473). The deduced polypeptide was found to share 53% amino acid sequence identity with Arabidopsis CBF1 and contain the PKK/RPAGR<sub>x</sub>KFxETRHP and DSAWR signature sequences (Fig. 1). Moreover, CBF-like transcripts were found to accumulate rapidly upon exposure of tomato plants to low temperature (Fig. 2). The results shown are from an experiment using tomato var. Castle Mart, but similar results were obtained with Bonny Best, Micro-Tom, and D Huang (not shown). Unlike in Arabidopsis, *B. napus*, rye, and wheat, however, the transcript levels of the tomato CBF transcripts in Castle Mart (Fig. 2) and the other varieties (not shown) appeared to return to those found in warm-grown plants after 24 h of exposure to low temperature and remained at low levels after 1 week of cold treatment (not shown). We were unable to test for the expression of tomato cold-regulated genes containing active CRT/DRE-like elements because to our knowledge, such genes have not yet been identified.

#### DISCUSSION

Cold acclimation in Arabidopsis involves action of the CBF cold-response pathway (Thomashow, 2001). The hallmark characteristics of this pathway are rapid induction of the CBF genes in response to low temperature followed by expression of the CBF regulon, which includes genes that increase plant freezing tolerance. Here, we report that *B. napus* encodes CBF-like proteins, that transcripts encoding these proteins accumulate rapidly in response to low temperature, and that this is closely followed by induction of *Bn115*, an ortholog of the CBF-targeted Arabidopsis gene *COR15a*. Moreover, we demonstrate that overexpression of Arabidopsis CBF genes in *B. napus* induces expression of *Bn115* and *Bn28*, an ortholog of the CBF-targeted Arabidopsis gene *COR6.6*, and increases freezing tolerance in both nonacclimated and cold-acclimated plants. From these results we conclude that *B. napus*, a close relative of Arabidopsis that cold acclimates, encodes a CBF cold-response pathway related to that found in Arabidopsis. In addition, we conclude that components of the CBF cold-response pathway are conserved in wheat and rye, more distant relatives of Arabidopsis that also cold acclimate. In particular, we show that these cereals encode CBF-like proteins, that transcripts for these proteins accumulate rapidly in response to low temperature and that this is quickly followed by induction of *Wcs120/COR39*, a gene with a cold-inducible promoter that has multiple copies of the CRT/DRE core sequence, CCGAC (Ouellet et al., 1998).

It is significant that the results presented also indicate that cold-regulated CBF-like genes are not limited to plants that cold acclimate. To be specific, we show that transcripts encoding a CBF-like protein(s) rapidly accumulate in response to low temperature in tomato, a chilling-sensitive plant that does not cold acclimate. Thus, tomato appears to have components of a CBF cold-response pathway. Thus, a fundamental question raised is why doesn't tomato cold acclimate? One possibility is that tomato has a completely functional CBF cold-response pathway, but that some other component(s) of the cold acclimation response is limiting. In an alternate manner, tomato might not have a fully functional CBF cold-response pathway. There might, for instance, be differences in the activities of the CBF-like proteins, though we have found that overexpression of the tomato CBF coding sequence (accession no. AY034473) in transgenic Arabidopsis plants activates expression of *COR15a* and *COR6.6* without a low temperature stimulus (X. Zhang and M.F. Thomashow, unpublished data). Other possibilities would include differences in the composition of the CBF regulons and differences in regulation of the CBF genes. The results presented indicate that the levels of the tomato CBF transcripts do not remain elevated at low temperature as Arabidopsis CBF transcripts do (Fig. 2). If true, it may be that an inability of tomato to sustain CBF expression results in only transient ex-

pression of CBF-targeted genes, which in turn may not allow the development of freezing (and possibly chilling) tolerance.

The AP2/EREBP protein family is characterized by a DNA-binding motif that is unique to plants, the AP2/EREBP domain (Riechmann and Meyerowitz, 1998). The domain consists of an  $\alpha$ -helix and a three-stranded antiparallel  $\beta$ -sheet that interacts with base pairs within the DNA major groove (Allen et al., 1998). The AP2/EREBP domain is found in a large number of plant proteins including more than 140 proteins in Arabidopsis (Riechmann et al., 2000). The results presented here indicate that the Arabidopsis CBF1, CBF2, and CBF3 proteins form a subset of the AP2/EREBP proteins that is characterized by two additional sequences that immediately surround the AP2/EREBP domain, PKK/RPAGRxKFxETRHP upstream of the domain and DSAWR downstream of it (Fig. 1). These "signature sequences" are present in CBF-like proteins from *B. napus*, wheat, rye, and tomato (Fig. 1). Conservation of these sequences across evolutionarily diverse plant species suggests that they have an important functional role. The resemblance of the PKK/RPAGRxKFxETRHP sequence to nuclear transport signals (Smith and Raikhel, 1999) indicates that it might be involved in protein trafficking as previously suggested (Stockinger et al., 1997). The signature sequences would not appear to be involved in recognition of the CRT/DRE regulatory element because they (or closely related sequences) are not present in the Arabidopsis AP2/EREBP protein DREB2a (Liu et al., 1998). This protein has been demonstrated to bind to the CRT/DRE element and activate gene expression in Arabidopsis in a transient assay (though interestingly not in stable Arabidopsis transformants; Liu et al., 1998). The *DREB2a* gene is not induced by low temperature, but instead is induced in response to dehydration stress (Liu et al., 1998). Expression of the DREB2a protein in drought-stressed plants is proposed to account, at least in part, for the dehydration responsiveness of the CRT/DRE element (Liu et al., 1998).

Understanding the mechanisms that plants have evolved to tolerate environmental stresses has the potential to provide new tools and strategies to improve the environmental stress tolerance of plants. The discovery of the Arabidopsis CBF cold-response pathway has possibilities in this regard. Previous studies demonstrated that increased expression of the CBF genes in Arabidopsis results in an increase in both freezing and drought tolerance (Jaglo-Ottosen et al., 1998; Liu et al., 1998; Kasuga et al., 1999; Gilmour et al., 2000). Here, we extend these findings to an important agronomic crop plant, *Brassica* oilseed rape (canola). We show that the freezing tolerance of *B. napus* can be enhanced through CBF-mediated "regulon engineering." It is important to bear in mind, however, that constitutive high-level overexpression of the CBF genes can result in undesirable agronomic

traits. In Arabidopsis, high-level CBF overexpression can cause a "stunted" growth phenotype, a decrease in seed yield and a delay in flowering (Liu et al., 1998; Gilmour et al., 2000). The CBF-expressing *B. napus* plants used in the experiments described here were grown in environmental chambers under constant light and did not exhibit overt adverse effects in growth and development, but when grown under greenhouse conditions, display stunted growth and delayed flowering phenotypes (V. Haake and J. Zhang, unpublished data). Whether strategies such as using stress-inducible promoters to drive CBF expression (Kasuga et al., 1999) can be developed to attain the potential positive effects of CBF regulon engineering without incurring undesirable negative traits remains to be determined.

## MATERIALS AND METHODS

### Plant Material

*Brassica napus* cv Westar (a spring-type canola), winter wheat (*Triticum aestivum* L. cv Norstar), winter rye (*Secale cereale* L. cv Puma), and tomato (*Lycopersicon esculentum* var. Bonny Best, Castle Mart, Micro-Tom, and D Huang) were grown in pots containing Baccto Planting Mix (Michigan Peat, Houston) in controlled environment chambers at 20°C to 22°C under continuous cool-white fluorescent illumination of 100 to 150  $\mu\text{mol m}^{-2} \text{s}^{-1}$  light intensity as described by Gilmour et al. (1988). For cold acclimation, plants were incubated at 4°C under continuous cool-white fluorescent illumination at approximately 50  $\mu\text{mol m}^{-2} \text{s}^{-1}$  light intensity.

### Isolation of cDNAs Encoding CBF-Like Proteins

A *B. napus* genomic DNA fragment encoding a CBF-like polypeptide was isolated by PCR (Innis et al., 1990) using degenerate primers O368 (CAYCCNATHAYMGNG-GNGT) and O378 (GGNARNARCATNCCYTCNGCC) based on conserved regions of the Arabidopsis CBF proteins at the beginning of the AP2/EREBP domain and putative activation domain, respectively. Full-length cDNAs were isolated based on the partial gene sequence using 5' and 3' RACE (MarathonTM cDNA amplification kit, CLONTECH, Palo Alto, CA). The isolation of cDNAs for rye and wheat CBF-like proteins was based on the sequence for a putative rice CBF1 homolog present in the GenBank EST database (accession no. AB023482). The rice gene was isolated from genomic DNA by PCR using primers O18016 (acgcgtcgac-CCATCATCACCGAGATCGACTCGAC) and O18017 (ataa-gaatcgcgccgTCATTGTTGCTCACTGGGAG). Based on the rice sequence, primers O18065 (GGCCGGCGGGGGC-GAACCAAGTTCC) and O18066 (AGGCAGAGTCGGCG-AAGTTGAGGC) were synthesized and PCR used to isolate CBF gene fragments from rye cDNA libraries of RNA prepared from cold-acclimated plants (J. Zhang and V. Haake, unpublished data). cDNAs encoding full-length rye CBF-like proteins were isolated by screening cDNA libraries using the cloned partial genes as probes. The wheat cDNA was isolated by screening a cDNA library (Guo et al., 1992) with one of the rye cDNAs (accession no. AF370730). A



tomato EST encoding a CBF-like protein (accession no. AI484513) was obtained from the Clemson University Genomics Institute (Clemson, SC). The sequence for the entire cDNA insert was determined and deposited (accession no. AY034473).

### Transformation of *B. napus*

The coding sequences for Arabidopsis CBF1, CBF2, and CBF3 were placed under control of the strong constitutive cauliflower mosaic virus 35S promoter in the plant expression vector pGA643 (An, 1995) which includes the NPTII gene to select for kanamycin resistance. The vector, with and without inserts, was introduced into *Agrobacterium tumefaciens* strain GV3101 and used to transform *B. napus* cotyledonary petioles selecting for kanamycin resistance (Moloney et al., 1989). Regenerated plants were tested for T-DNA inserts using an NPTII ELISA kit (5 Prime-3 Prime, Inc., Boulder, CO). Positive T<sub>0</sub> plants were self-pollinated and T<sub>1</sub> seeds collected. Because T<sub>1</sub> populations were not homozygous for T-DNA inserts, individual plants were tested either for expression of the NPTII gene using the NPTII ELISA assay or for the presence of the NPTII gene using the PCR (primers were 5': TGGAGAGGCTATTCGCTA and 3': CACCATGATATTCGGCAAG) before being used in experiments.

### RNA Hybridization

Total RNA was isolated from *B. napus* using TRIZOL reagent (GibcoBRL, Grand Island, NY), from wheat and rye plants using a Plant RNA Isolation Kit (Qiagen Inc., Valencia, CA), and from tomato (Howe et al., 1996) and Arabidopsis (Gilmour et al., 2000) as described. Northern transfers (5–20 µg total RNA) were prepared, hybridized, and washed as described (Stockinger et al., 1997). The probe for Arabidopsis CBF1 was prepared from a full-length cDNA of CBF1 (Stockinger et al., 1997; Gilmour et al., 2000). The probe for *B. napus* CBF transcripts was made by PCR amplification of genomic DNA using 5' and 3' primers, GGT-TACGTTAGCGGAGAGT and GGACGGCGGCGGCAAAAG, respectively, based on sequence AF084185. The probe for rye and wheat CBF transcripts was the entire insert from one of the cloned rye cDNAs (accession no. AF370730). The probe for tomato CBF transcripts was the entire cDNA insert from EST AI484513. Hybridization probes for BN28 (Orr et al., 1992) and BN115 (Weretilnyk et al., 1993) were the entire cDNA inserts in plasmids pBN28 and pBN115, respectively, kindly provided by Jas Singh (Agriculture Canada, Ottawa). The probe for wheat COR39 was the entire cDNA insert from pWG1 (Guo et al., 1992). DNA fragments were <sup>32</sup>P radiolabeled (Stockinger et al., 1997; Gilmour et al., 2000) and gel purified (Sambrook et al., 1989) as described.

### Immunoblot Analysis

Total protein was extracted by grinding frozen tissue (approximately 300 mg) in extraction buffer (approximately 300 µL) containing 50 mM Tris-HCl (pH 8.0), 5% (w/v) glycerol, 100 mM KCl, and 1.5% (w/v) polyvinyl-polypyr-

rolidone. Insoluble material was removed by centrifugation at 13,000g for 20 min at 4°C. Protein concentrations of supernatants were determined using the Bradford dye-binding assay (Bio-Rad, Hercules, CA). Total soluble protein (100 µg) was fractionated by 10% (w/v) acrylamide tricine SDS/PAGE (Schägger and von Jagow, 1987) and transferred to 0.1-µm nitrocellulose membranes by electroblotting (Towbin et al., 1979) as described (Artus et al., 1996). BN28 protein was detected using antiserum kindly provided by Anne Johnson (Boothe et al., 1997) and visualized using the enhanced chemiluminescence system (Amersham, Buckinghamshire, UK).

### Freezing Tolerance Assays

*B. napus* T<sub>1</sub> seedlings (approximately 2 weeks old) were screened for the presence of the transgene and thinned to one plant per pot. At 4 to 6 weeks, plants were either tested directly for freezing tolerance (nonacclimated plants) or were placed at 4°C under continuous fluorescent illumination of approximately 50 µmol m<sup>-2</sup> s<sup>-1</sup> for 3 weeks. Freezing tolerance was determined using the electrolyte leakage test as previously described (Jaglo-Ottosen et al., 1998; Gilmour et al., 2000). Tissue from the smallest two leaves was obtained using a 6-mm paper punch. Three or four punches were used in each of three replicate samples for each temperature point tested. The EL<sub>50</sub> values (temperature that caused leakage of 50% of the electrolytes) were determined by fitting model curves of up to third-order linear polynomials for each electrolyte leakage test. To ensure unbiased predictions of electrolyte leakage, trends significantly improving the model fit at the 0.2 probability level were retained. An unbalanced one-way analysis of variance (ANOVA), adjusted for the different number of EL<sub>50</sub> values for each tissue type was determined using SAS PROC GLM (SAS Institute, 1989).

### ACKNOWLEDGMENTS

We are grateful to Wilf Keller for hosting one of us (S.K.) in his laboratory to learn how to transform canola; Maurice Moloney for advice regarding canola transformation; Jas Singh for cDNAs encoding BN28 and BN115; Anne Johnson for the antibody to the BN28 protein; Trevor Wagner for conducting initial alignments of CBF proteins; Cai-Zhong Jiang, Mark Leibman, and Sanjeev Pillai for their help in isolating CBF homologs; and Steve Triezenberg and Sarah Gilmour for critical reading of the manuscript.

Received June 21, 2001; returned for revision July 17, 2001; accepted August 7, 2001.

### LITERATURE CITED

- Allen MD, Yamasaki K, Ohme-Takagi M, Tateno M, Suzuki M (1998) A novel mode of DNA recognition by a beta-sheet revealed by the solution structure of the GCC-box binding domain in complex with DNA. *EMBO J* 17: 5484–5496
- An G (1995) Binary Ti plasmid vectors. *Methods Mol Biol* 44: 47–58
- Artus NN, Uemura M, Steponkus PL, Gilmour SJ, Lin CT, Thomashow MF (1996) Constitutive expression of the

- cold-regulated *Arabidopsis thaliana* COR15a gene affects both chloroplast and protoplast freezing tolerance. *Proc Natl Acad Sci USA* 93: 13404–13409
- Baker SS, Wilhelm KS, Thomashow MF (1994) The 5'-region of *Arabidopsis thaliana* cor15a has cis-acting elements that confer cold-, drought- and ABA-regulated gene expression. *Plant Mol Biol* 24: 701–713
- Boothe JG, Sonnichsen FD, de Beus MD, Johnson-Flanagan AM (1997) Purification, characterization, and structural analysis of a plant low-temperature-induced protein. *Plant Physiol* 113: 367–376
- Gilmour SJ, Artus NN, Thomashow MF (1992) cDNA sequence analysis and expression of two cold-regulated genes of *Arabidopsis thaliana*. *Plant Mol Biol* 18: 13–21
- Gilmour SJ, Hajela RK, Thomashow MF (1988) Cold acclimation in *Arabidopsis thaliana*. *Plant Physiol* 87: 745–750
- Gilmour SJ, Sebolt AM, Salazar MP, Everard JD, Thomashow MF (2000) Overexpression of the *Arabidopsis* CBF3 transcriptional activator mimics multiple biochemical changes associated with cold acclimation. *Plant Physiol* 124: 1854–1865
- Gilmour SJ, Zarka DG, Stockinger EJ, Salazar MP, Houghton JM, Thomashow MF (1998) Low temperature regulation of the *Arabidopsis* CBF family of AP2 transcriptional activators as an early step in cold-induced COR gene expression. *Plant J* 16: 433–442
- Guo W, Ward RW, Thomashow MF (1992) Characterization of a cold-regulated wheat gene related to *Arabidopsis* cor47. *Plant Physiol* 100: 915–922
- Hajela RK, Horvath DP, Gilmour SJ, Thomashow MF (1990) Molecular cloning and expression of cor (cold-regulated) genes in *Arabidopsis thaliana*. *Plant Physiol* 93: 1246–1252
- Houde M, Danyluk J, Laliberte JF, Rassart E, Dhindsa RS, Sarhan F (1992) Cloning, characterization, and expression of a cDNA encoding a 50-kilodalton protein specifically induced by cold acclimation in wheat. *Plant Physiol* 99: 1381–1387
- Howe GA, Lightner J, Browse J, Ryan CA (1996) An octadecanoid pathway mutant (JL5) of tomato is compromised in signaling for defense against insect attack. *Plant Cell* 8: 2067–2077
- Hughes MA, Dunn MA (1996) The molecular biology of plant acclimation to low temperature. *J Exp Bot* 47: 291–305
- Innis MA, Gelfand GD, Sninsky JJ, White TJ (1990) PCR Protocols: A Guide to Methods and Applications. Academic Press Inc., San Diego
- Jaglo-Ottosen KR, Gilmour SJ, Zarka DG, Schabenberger O, Thomashow MF (1998) *Arabidopsis* CBF1 overexpression induces COR genes and enhances freezing tolerance. *Science* 280: 104–106
- Jiang C, Iu B, Singh J (1996) Requirement of a CCGAC cis-acting element for cold induction of the BN115 gene from winter *Brassica napus*. *Plant Mol Biol* 30: 679–684
- Kasuga M, Liu Q, Miura S, Yamaguchi-Shinozaki K, Shinozaki K (1999) Improving plant drought, salt, and freezing tolerance by gene transfer of a single stress-inducible transcription factor. *Nat Biotechnol* 17: 287–291
- Liu Q, Kasuga M, Sakuma Y, Abe H, Miura S, Yamaguchi-Shinozaki K, Shinozaki K (1998) Two transcription factors, DREB1 and DREB2, with an EREBP/AP2 DNA binding domain separate two cellular signal transduction pathways in drought- and low-temperature-responsive gene expression, respectively, in *Arabidopsis*. *Plant Cell* 10: 1391–1406
- Moloney M, Walker JM, Sharma KK (1989) High-efficiency transformation of *Brassica-napus* using *Agrobacterium* vectors. *Plant Cell Rep* 8: 238–242
- Orr W, Iu B, White TC, Robert LS, Singh J (1992) Complementary DNA sequence of a low temperature-induced *Brassica napus* gene with homology to the *Arabidopsis thaliana* kin1 gene. *Plant Physiol* 98: 1532–1534
- Ouellet F, Vazquez-Tello A, Sarhan F (1998) The wheat wcs120 promoter is cold-inducible in both monocotyledonous and dicotyledonous species. *FEBS Lett* 423: 324–328
- Riechmann JL, Heard J, Martin G, Reuber L, Jiang C, Keddie J, Adam L, Pineda O, Ratcliffe OJ, Samaha RR et al. (2000) *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. *Science* 290: 2105–2110
- Riechmann JL, Meyerowitz EM (1998) The AP2/EREBP family of plant transcription factors. *Biol Chem* 379: 633–646
- Sakai A, Larcher W (1987) Frost Survival of Plants: Responses and Adaptation to Freezing Stress. Springer-Verlag, Berlin
- Sambrook J, Fritsch E, Maniatis T (1989) Molecular Cloning: A Laboratory Manual, Ed 2. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
- SAS Institute (1989) SAS/STAT User's Guide, version 6. SAS Institute, Cary, NC
- Schägger H, von Jagow G (1987) Tricine-sodium dodecyl sulfate-polyacrylamide gel electrophoresis for the separation of proteins in the range from 1 to 100 kDa. *Anal Biochem* 166: 368–379
- Smith HM, Raikhel NV (1999) Protein targeting to the nuclear pore: what can we learn from plants? *Plant Physiol* 119: 1157–1164
- Stockinger EJ, Gilmour SJ, Thomashow MF (1997) *Arabidopsis thaliana* CBF1 encodes an AP2 domain-containing transcriptional activator that binds to the C-repeat/DRE, a cis-acting DNA regulatory element that stimulates transcription in response to low temperature and water deficit. *Proc Natl Acad Sci USA* 94: 1035–1040
- Thomashow MF (1999) Plant cold acclimation: freezing tolerance genes and regulatory mechanisms. *Annu Rev Plant Physiol Plant Mol Biol* 50: 571–599
- Thomashow MF (2001) So what's new in the field of plant cold acclimation? Lots! *Plant Physiol* 125: 89–93
- Towbin H, Staehelin T, Gordon J (1979) Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets; procedure and some applications. *Proc Natl Acad Sci USA* 76: 4350–4354
- Weretilnyk E, Orr W, White TC, Iu B, Singh J (1993) Characterization of three related low-temperature-regulated cDNAs from winter *Brassica napus*. *Plant Physiol* 101: 171–177
- Yamaguchi-Shinozaki K, Shinozaki K (1994) A novel cis-acting element in an *Arabidopsis* gene is involved in responsiveness to drought, low-temperature, or high-salt stress. *Plant Cell* 6: 251–264



# Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis

Jonathan A. Eisen<sup>1</sup>

Department of Biological Sciences, Stanford University, Stanford, California 94305-5020 USA

**T**he ability to accurately predict gene function based on gene sequence is an important tool in many areas of biological research. Such predictions have become particularly important in the genomics age in which numerous gene sequences are generated with little or no accompanying experimentally determined functional information. Almost all functional prediction methods rely on the identification, characterization, and quantification of sequence similarity between the gene of interest and genes for which functional information is available. Because sequence is the prime determining factor of function, sequence similarity is taken to imply similarity of function. There is no doubt that this assumption is valid in most cases. However, sequence similarity does not ensure identical functions, and it is common for groups of genes that are similar in sequence to have diverse (although usually related) functions. Therefore, the identification of sequence similarity is frequently not enough to assign a predicted function to an uncharacterized gene; one must have a method of choosing among similar genes with different functions. In such cases, most functional prediction methods assign likely functions by quantifying the levels of similarity among genes. I suggest that functional predictions can be greatly improved by focusing on *how* the genes became similar in sequence (i.e., evolution) rather than on the sequence similarity itself. It is well established that many aspects of comparative biology can benefit from evolutionary studies (Felsenstein 1985), and comparative molecular biology is no exception

(e.g., Altschul et al. 1989; Goldman et al. 1996). In this commentary, I discuss the use of evolutionary information in the prediction of gene function. To appreciate the potential of a *phylogenomic* approach to the prediction of gene function, it is necessary to first discuss how gene sequence is commonly used to predict gene function and some general features about gene evolution.

## Sequence Similarity, Homology, and Functional Predictions

To make use of the identification of sequence similarity between genes, it is helpful to understand how such similarity arises. Genes can become similar in sequence either as a result of *convergence* (similarities that have arisen without a common evolutionary history) or descent with modification from a common ancestor (also known as *homology*). It is imperative to recognize that sequence similarity and homology are not interchangeable terms. Not all homologs are similar in sequence (i.e., homologous genes can diverge so much that similarities are difficult or impossible to detect) and not all similarities are due to homology (Reeck et al. 1987; Hillis 1994). Similarity due to convergence, which is likely limited to small regions of genes, can be useful for some functional predictions (Henikoff et al. 1997). However, most sequence-based functional predictions are based on the identification (and subsequent analysis) of similarities that are thought to be due to homology. Because homology is a statement about common ancestry, it cannot be proven directly from sequence similarity. In these cases, the inference of homology is made based on finding levels of sequence similarity that are thought to be too high to be due to

convergence (the exact threshold for such an inference is not well established).

Improvements in database search programs have made the identification of likely homologs much faster, easier, and more reliable (Altschul et al. 1997; Henikoff et al. 1998). However, as discussed above, in many cases the identification of homologs is not sufficient to make specific functional predictions because not all homologs have the same function. The available similarity-based functional prediction methods can be distinguished by how they choose the homolog whose function is most relevant to a particular uncharacterized gene (Table 1). Some methods are relatively simple—many researchers use the highest scoring homolog (as determined by programs like BLAST or BLAZE) as the basis for assigning function. While highest hit methods are very fast, can be automated readily, and are likely accurate in many instances, they do not take advantage of any information about how genes and gene functions evolve. For example, gene duplication and subsequent divergence of function of the duplicates can result in homologs with different functions being present within one species. Specific terms have been created to distinguish homologs in these cases (Table 2): Genes of the same duplicate group are called *orthologs* (e.g.,  $\beta$ -globin from mouse and humans), and different duplicates are called *paralogs* (e.g.,  $\alpha$ - and  $\beta$ -globin) (Fitch 1970). Because gene duplications are frequently accompanied by functional divergence, dividing genes into groups of orthologs and paralogs can improve the accuracy of functional predictions. Recognizing that the one-to-one sequence comparisons used by most methods do not reliably distinguish orthologs from paralogs, Tatusov et al. (1997) developed the COG cluster-

<sup>1</sup>E-MAIL [jeisen@leland.stanford.edu](mailto:jeisen@leland.stanford.edu); FAX (650) 725-1848.  
WWW: <http://www.leland.stanford.edu/~jeisen>.

**Table 1. Methods of Predicting Gene Function When Homologs Have Multiple Functions**

## Highest Hit

The uncharacterized gene is assigned the function (or frequently, the annotated function) of the gene that is identified as the highest hit by a similarity search program (e.g., Tomb et al. 1997).

## Top Hits

Identify top 10+ hits for the uncharacterized gene. Depending on the degree of consensus of the functions of the top hits, the query sequence is assigned a specific function, a general activity with unknown specificity, or no function (e.g., Blattner et al. 1997).

## Clusters of Orthologous Groups

Genes are divided into groups of orthologs based on a cluster analysis of pairwise similarity scores between genes from different species. Uncharacterized genes are assigned the function of characterized orthologs (Tatusov et al. 1997).

## Phylogenomics

Known functions are overlaid onto an evolutionary tree of all homologs. Functions of uncharacterized genes are predicted by their phylogenetic position relative to characterized genes (e.g., Eisen et al. 1995, 1997).

ing method (see Table 1). Although the COG method is clearly a major advance in identifying orthologous groups of genes, it is limited in its power because clustering is a way of classifying levels of similarity and is not an accurate method of inferring evolutionary relationships (Swofford et al. 1996). Thus, as sequence similarity and clustering are not reliable estimators of evolutionary relatedness, and as the incorporation of such phylogenetic information has been so useful to other areas of biology, evolutionary techniques should be useful for improving the accuracy of predicting function based on sequence similarity.

## Phylogenomics

There are many ways in which evolu-

tionary information can be used to improve functional predictions. Below, I present an outline of one such *phylogenomic* method (see Fig. 1), and I compare this method to nonevolutionary functional prediction methods. This method is based on a relatively simple assumption—because gene functions change as a result of evolution, reconstructing the evolutionary history of genes should help predict the functions of uncharacterized genes. The first step is the generation of a phylogenetic tree representing the evolutionary history of the gene of interest and its homologs. Such trees are distinct from clusters and other means of characterizing sequence similarity because they are inferred by special techniques that help convert patterns of similarity into evolutionary relationships (see Swofford et al. 1996). After the gene tree is inferred, biologically determined functions of the various homologs are overlaid onto the tree. Finally, the structure of the tree and the relative phylogenetic positions of genes of different functions are used to trace the history of functional changes, which is then used to predict functions of uncharacterized genes. More detail of this method is provided below.

## Identification of Homologs

The first step in studying the evolution of a particular gene is the identification of homologs. As with similarity-based functional prediction methods, likely homologs of a particular gene are identified through database searches. Because phylogenetic methods benefit greatly from more data, it is useful to augment this initial list by using identified homologs as queries for further

database searches or using automatic iterated search methods such as PSI-BLAST (Altschul et al. 1997). If a gene family is very large (e.g., ABC transporters), it may be necessary to only analyze a subset of homologs. However, this must be done with extreme care, as one might accidentally leave out proteins that would be important for the analysis.

## Alignment and Masking

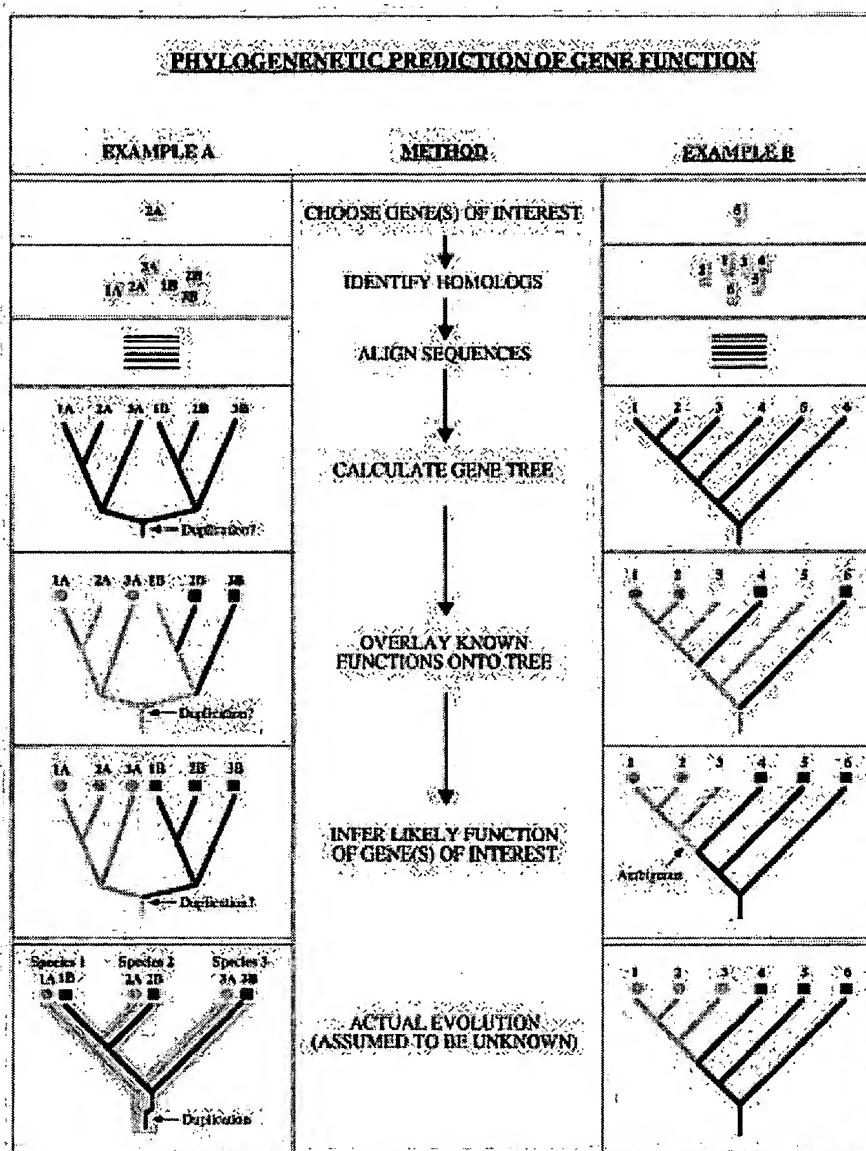
Sequence alignment for phylogenetic analysis has a particular purpose—it is the assignment of *positional* homology. Each column in a multiple sequence alignment is assumed to include amino acids or nucleotides that have a common evolutionary history, and each column is treated separately in the phylogenetic analysis. Therefore, regions in which the assignment of positional homology is ambiguous should be excluded (Gatesy et al. 1993). The exclusion of certain alignment positions (also known as masking) helps to give phylogenetic methods much of their discriminatory power. Phylogenetic trees generated without masking (as is done in many sequence analysis software packages) are less likely to accurately reflect the evolution of the genes than trees with masking.

## Phylogenetic Trees

For extensive information about generating phylogenetic trees from sequence alignments, see Swofford et al. (1996). In summary, there are three methods commonly used: parsimony, distance, and maximum likelihood (Table 3), and each has its advantages and disadvantages. I

**Table 2. Types of Molecular Homology**

|                     |   |
|---------------------|---|
| Homolog             | Genes that are descended from a common ancestor (e.g., all globins)   |
| Ortholog            | Homologous genes that have diverged from each other after <i>speciation</i> events (e.g., human $\beta$ - and chimp $\beta$ -globin)          |
| Paralog             | Homologous genes that have diverged from each other after <i>gene duplication</i> events (e.g., $\beta$ - and $\gamma$ -globin)               |
| Xenolog             | Homologous genes that have diverged from each other after <i>lateral gene transfer</i> events (e.g., antibiotic resistance genes in bacteria) |
| Positional homology | Common ancestry of specific amino acid or nucleotide positions in different genes   |



**Figure 1** Outline of a phylogenomic methodology. In this method, information about the evolutionary relationships among genes is used to predict the functions of uncharacterized genes (see text for details). Two hypothetical scenarios are presented and the path of trying to infer the function of two uncharacterized genes in each case is traced. (A) A gene family has undergone a gene duplication that was accompanied by functional divergence. (B) Gene function has changed in one lineage. The true tree (which is assumed to be unknown) is shown at the bottom. The genes are referred to by numbers (which represent the species from which these genes come) and letters (which in A represent different genes within a species). The thin branches in the evolutionary trees correspond to the gene phylogeny and the thick gray branches in A (bottom) correspond to the phylogeny of the species in which the duplicate genes evolve in parallel (as paralogs). Different colors (and symbols) represent different gene functions; gray (with hatching) represents either unknown or unpredictable functions.

prefer distance methods because they are the quickest when using large data sets. Before using any particular tree it is important to estimate the robustness and accuracy of the phylogenetic pat-

terns it shows (through techniques such as the comparison of trees generated by different methods and bootstrapping). Finally, in most cases, it is also useful to determine a root for the tree.

## Functional Predictions

To make functional predictions based on the phylogenetic tree, it is necessary to first overlay any known functions onto the tree. There are many ways this "map" can then be used to make functional predictions, but I recommend splitting the task into two steps. First, the tree can be used to identify likely gene duplication events in the past. This allows the division of the genes into groups of orthologs and paralogs (e.g., Eisen et al. 1995). Uncharacterized genes can be assigned a likely function if the function of any ortholog is known (and if all characterized orthologs have the same function). Second, parsimony reconstruction techniques (Maddison and Maddison 1992) can be used to infer the likely functions of uncharacterized genes by identifying the evolutionary scenario that requires the fewest functional changes over time (Fig. 1). The incorporation of more realistic models of functional change (and not just minimizing the total number of changes) may prove to be useful, but the parsimony minimization methods are probably sufficient in most cases.

## Is the Phylogenomic Method Worth the Trouble?

Phylogenomic methods require many more steps and usually much more manual labor than similarity-based functional prediction methods. Is the phylogenomic approach worth the trouble? Many specific examples exist in which gene function has been shown to correlate well with gene phylogeny (Eisen et al. 1995; Atchley and Fitch 1997). Although no systematic comparisons of phylogenetic versus similarity-based functional prediction methods have been done, there are a variety of reasons to believe that the phylogenomic method should produce more accurate predictions than similarity-based methods. In particular, there are many conditions in which similarity-based methods are likely to make inaccurate predictions but which can be dealt with well by phylogenetic methods (see Table 4).

A specific example helps illustrate a potential problem with similarity-based methods. Molecular phylogenetic methods show conclusively that mycoplasmas share a common ancestor with low-GC Gram-positive bacteria (Weisburg et

# Insight/Outlook

**Table 3. Molecular Phylogenetic Methods**




| Method             |   |
|--------------------|---|
| Parsimony          | Possible trees are compared and each is given a score that is a reflection of the minimum number of character state changes (e.g., amino acid substitutions) that would be required over evolutionary time to fit the sequences into that tree. The optimal tree is considered to be the one requiring the fewest changes (the most parsimonious tree).   |
| Distance           | The optimal tree is generated by first calculating the estimated evolutionary distance between all pairs of sequences. Then these distances are used to generate a tree in which the branch patterns and lengths best represent the distance matrix.  |
| Maximum likelihood | Maximum likelihood is similar to parsimony methods in that possible trees are compared and given a score. The score is based on how likely the given sequences are to have evolved in a particular tree given a model of amino acid or nucleotide substitution probabilities. The optimal tree is considered to be the one that has the highest probability.  |
| Bootstrapping      | Alignment positions within the original multiple sequence alignment are resampled and new data sets are made. Each bootstrapped data set is used to generate a separate phylogenetic tree and the trees are compared. Each node of the tree can be given a bootstrap percentage indicating how frequently those species joined by that node group together in different trees. Bootstrap percentage does not correspond directly to a confidence limit. |

al. 1989). However, examination of the percent similarity between mycoplasmal genes and their homologs in bacteria does not clearly show this relationship.

This is because mycoplasmas have undergone an accelerated rate of molecular evolution relative to other bacteria. Thus, a BLAST search with a gene from

*Bacillus subtilis* (a low GC Gram-positive species) will result in a list in which the mycoplasma homologs (if they exist) score lower than genes from many spe-

**Table 4. Examples of Conditions in Which Similarity Methods Produce Inaccurate Predictions of Function**

| Evolutionary Pattern and Type of Similarity Prediction   | Gene With Putative Function <sup>1</sup> | Highest Hit Method              |                       | Phylogenetic Method            |                       | Comments   |
|--|--|---------------------------------|-----------------------|--------------------------------|-----------------------|--|
|  |  | Predicted Function <sup>2</sup> | Accuracy <sup>3</sup> | Putative Function <sup>4</sup> | Accuracy <sup>3</sup> |  |
| A. Functional change during evolution.<br>  | 1*                                       | ■                               | +                     | ■                              | +                     | <ul style="list-style-type: none"> <li>Phylogenetic method cannot predict functions for all genes, but the predictions that are made are accurate.</li> <li>Highest hit method is misleading because function changed among homologs but hierarchies of similarity do not correlate with the function (see Bolker and Ruff 1996).</li> </ul>   |
|  | 2*                                       | ■                               | +                     | ■                              | +                     |  |
|  | 3*                                       | ■                               | +                     | ■                              | +                     |  |
|  | 4*                                       | ■                               | +                     | ■                              | +                     |  |
|  | 5*                                       | ■                               | +                     | ■                              | +                     |  |
|  | 6*                                       | ■                               | +                     | ■                              | +                     |  |
| B. Functional change & rate variation.<br>  | 1*                                       | ■                               | +                     | ■                              | +                     | <ul style="list-style-type: none"> <li>Similarity based methods perform particularly poorly when evolutionary rates vary between taxa.</li> <li>Molecular phylogenetic methods can allow for rate variation and reconstruct gene history reasonably accurately.</li> </ul>   |
|  | 2*                                       | ■                               | +                     | ■                              | +                     |  |
|  | 3*                                       | ■                               | +                     | ■                              | +                     |  |
|  | 4*                                       | ■                               | +                     | ■                              | +                     |  |
|  | 5*                                       | ■                               | +                     | ■                              | +                     |  |
|  | 6*                                       | ■                               | +                     | ■                              | +                     |  |
| C. Gene duplication and rate variation.<br> | 1A*                                      | ■                               | +                     | ■                              | +                     | <ul style="list-style-type: none"> <li>Most similarity based methods are not ideally set up to deal with cases of gene duplication since orthologous genes do not always have significantly more sequence similarity to each other than to paralogues (Eisen et al. 1993; Zardoya et al. 1996; Tatusov et al. 1997).</li> <li>Similarity-based methods perform particularly poorly when rate variation and gene duplication are combined. This even applies to the COG method (see Table 1) since it works by classifying levels of similarity and not by inferring history. Nevertheless, the COG method is a significant improvement over other similarity based methods in classifying orthologs.</li> <li>Phylogenetic reconstruction is the most reliable way to infer gene duplication events and thus determine orthology.</li> </ul> |
|  | 2A*                                      | ■                               | +                     | ■                              | +                     |  |
|  | 3A*                                      | ■                               | +                     | ■                              | +                     |  |
|  | 1B*                                      | ■                               | +                     | ■                              | +                     |  |
|  | 2B*                                      | ■                               | +                     | ■                              | +                     |  |
|  | 3B*                                      | ■                               | +                     | ■                              | +                     |  |

\* The true tree is shown but it is assumed that it is not known. Different circles and symbols represent different functions. Numbers correspond to different species.

† The function of all other genes is assumed to be known.

‡ The top hit can be determined from the tree by finding the gene in the shortest evolutionary distance away (as determined along the branches of the tree).

§ It is assumed that the tree of the genes can be reconstructed accurately by molecular phylogenetic methods (see Fig. 1).

cies of bacteria less closely related to *B. subtilis*. When amounts or rates of change vary between lineages, phylogenetic methods are better able to infer evolutionary relationships than similarity methods (including clustering) because they allow for evolutionary branches to have different lengths. Thus, in those cases in which gene function correlates with gene phylogeny and in which amounts or rates of change vary between lineages, similarity-based methods will be more likely than phylogenomic methods to make inaccurate functional predictions (see Table 4).

Another major advantage of phylogenetic methods over most similarity methods comes from the process of masking (see above). For example, a deletion of a large section of a gene in one species will greatly affect similarity measures but may not affect the function of that gene. A phylogenetic analysis including these genes could exclude the region of the deletion from the analysis by masking. In addition, regions of genes that are highly variable between species are more likely to undergo convergence and such regions can be excluded from phylogenetic analysis by masking. Masking thus allows the exclusion of regions of genes in which sequence similarity is likely to be "noisy" or misleading rather than a biologically important signal. The pairwise sequence comparisons used by most similarity-based functional prediction methods do not allow such masking. Phylogenetic methods have been criticized because of their dependence (for most methods) on multiple sequence alignments that are not always reliable and unbiased. However, multiple sequence alignments also allow for masking, which is probably more valuable than the cost of depending on alignments.

The conditions described above and highlighted in Table 4 are just some examples of conditions in which evolutionary methods are more likely to make accurate functional predictions than similarity-based methods. Phylogenetic methods are particularly useful when the history of a gene family includes many of these conditions (e.g., multiple gene duplications plus rate variation) or when the gene family is very large. The principle is simple—the more complicated the history of a gene family, the more useful it is to try to infer that history. Thus although the phylogenomic

method is slow and labor intensive, I believe it is worth using if accuracy is the main objective. In addition, information about the evolutionary relationships among gene homologs is useful for summarizing relationships among genes and for putting functional information into a useful context.

Despite the evolution of these methods, and likely continued improvements in functional predictions, it must be remembered that the key word is *prediction*. All methods are going to make inaccurate predictions of functions. For example, none of the methods described can perform well when gene functions can change with little sequence change as has been seen in proteins like opsins (Yokoyama 1997). Thus, sequence databases and genome researchers should make clear which functions assigned to genes are based on predictions and which are based on experiments. In addition, all prediction methods should use only experimentally determined functions as their grist for predictions. This will hopefully limit error propagation that can happen by using an inaccurate prediction of function to then predict the function of a new gene, which is a particular problem for the highest hit methods, as they rely on the function of only one gene at a time to make predictions (Eisen et al. 1997). Despite these and other potential problems, functional predictions are of great value in guiding research and in sorting through huge amounts of data. I believe that the increased use of phylogenetic methods can only serve to improve the accuracy of such functional predictions.

## REFERENCES

- Altschul, S.F., R.J. Carroll, and D.J. Lipman. 1989. *J. Mol. Biol.* **207**: 647–653.
- Altschul, S.F., T.L. Madden, A.A. Schaeffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. *Nucleic Acids Res.* **25**: 3389–3402.
- Atchley, W.R. and W.M. Fitch. 1997. *Proc. Natl. Acad. Sci.* **94**: 5172–5176.
- Blattner, F.R., G.I. Plunkett, C.A. Bloch, N.T. Perna, V. Burland, M. Riley, J. Collado-Vides, J.D. Glasner, C.K. Rode, G.F. Mayhew et al. 1997. *Science* **277**: 1453–1462.
- Bolker, J.A. and R.A. Raff. 1996. *BioEssays* **18**: 489–494.
- Eisen, J.A., D. Kaiser, and R.M. Myers. 1997. *Nature (Med.)* **3**: 1076–1078.
- Eisen, J.A., K.S. Sweder, and P.C. Hanawalt. 1995. *Nucleic Acids Res.* **23**: 2715–2723.
- Felsenstein, J. 1985. *Am. Nat.* **125**: 1–15.
- Fitch, W.M. 1970. *Syst. Zool.* **19**: 99–113.
- Gatesy, J., R. Desalle, and W. Wheller. 1993. *Mol. Phylog. Evol.* **2**: 152–157.
- Goldman, N., J.L. Thorne, and D.T. Jones. 1996. *J. Mol. Biol.* **263**: 196–208.
- Henikoff, S., E.A. Greene, S. Pietrovsky, P. Bork, T.K. Attwood, and L. Hood. 1997. *Science* **278**: 609–614.
- Henikoff, S., S. Pietrovski, and J.G. Henikoff. 1998. *Nucleic Acids Res.* **26**: 311–315.
- Hillis, D.M. 1994. In *Homology: The hierarchical basis of comparative biology* (ed. B.K. Hall), pp. 339–368. Academic Press, San Diego, CA.
- Maddison, W.P. and D.R. Maddison. 1992. *MacClade*. Sinauer Associates, Sunderland, MA.
- Reeck, G.R., C. Haën, D.C. Teller, R.F. Doolittle, W.M. Fitch, R.E. Dickerson, P. Chambon, A.D. McLachlan, E. Margoliash, T.H. Jukes et al. 1987. *Cell* **50**: 667.
- Swofford, D.L., G.J. Olsen, P.J. Waddell, and D.M. Hillis. 1996. In *Molecular systematics* (ed. D.M. Hillis, C. Moritz, and B.K. Mable), pp. 407–514. Sinauer Associates, Sunderland, MA.
- Tatusov, R.L., E.V. Koonin, and D.J. Lipman. 1997. *Science* **278**: 631–637.
- Tomb, J.F., O. White, A.R. Kerlavage, R.A. Clayton, G.G. Sutton, R.D. Fleischmann, K.A. Ketchum, H.P. Klenk, S. Gill, B.A. Dougherty et al. 1997. *Nature* **388**: 539–547.
- Weisburg, W.G., J.G. Tully, D.L. Rose, J.P. Petzel, H. Oyaizu, D. Yang, L. Mandelco, J. Sechrest, T.G. Lawrence, J. Van Etten et al. 1989. *J. Bacteriol.* **171**: 6455–6467.
- Yokoyama, S. 1997. *Annu. Rev. Genet.* **31**: 315–336.
- Zardoya, R., E. Abouheif, and A. Meyer. 1996. *Trends Genet.* **2**: 496–497.

# Structural genomics and signaling domains

**James H. Hurley, D. Eric Anderson, Bridgette Beach, Bertram Canagarajah, Yew Seng Jonathan Ho, Eudora Jones, Greg Miller, Saurav Misra, Matt Pearson, Layla Saidi, Silke Suer, Ray Trievel and Yosuke Tsujishita**

**Many novel signal transduction domains are being identified in the wake of genome sequencing projects and improved sensitivity in homology-detection techniques. The functions of these domains are being discovered by hypothesis-driven experiments and structural genomics approaches. This article reviews the recent highlights of research on modular signaling domains, and the relative contributions and limitations of the various approaches being used.**

The concept of modular signaling domains has been at the center of signal transduction research for the past 15 or so years. The discovery of domain families such as SH2, SH3, PDZ, PH, C1 and C2, and the understanding of their properties, have contributed immeasurably to our knowledge of signaling pathways. The importance of the domain concept is reflected in the fact that each one of the domains mentioned above occurs in hundreds of different signaling proteins. Once the function of a particular domain from one protein is well understood, powerful and testable inferences can be made as to the function of the many other proteins that contain that domain. Thus, domain information provides the simplest and most powerful conceptual bridge between otherwise overwhelmingly vast and complex sequence data. Because of this, great effort has gone into understanding the structures and functions of these domains, leading to our ability to rationalize, and in some cases predict, subtle differences in specificity.

The number of known signaling domain families has expanded rapidly in the past few years. This trend is driven by the value of the biological information to be gleaned, and is made possible by data from genome sequencing and by high-sensitivity sequence-homology detection. In 1997, all the known intracellular signaling domains were described in the pages of *TiBS* [1], a total of 37 domains. The signaling domain database SMART (Ref. [2] and <http://smart.embl-heidelberg.de/>) now contains 150 entries categorized as 'intracellular signaling'. Newly discovered domain families mined from homology searches of worldwide sequence databases are similar to raw ore in that the most valuable content must still be 'smelted' out of these sequences by other techniques. Bioinformatics-based discoveries of new signaling domain families sometimes define biochemical function in a clear-cut

way, as when one or more family members correspond to protein fragments whose activity has been previously characterized. At least as often, the identification of a new protein family poses the new question: what is (are) the function(s) of the FITB (fill in the blank) domain? The volume of domain discoveries is so great, and this question is raised so frequently, that it has spawned a new enterprise dedicated to its answer.

The problem of identifying the function of a protein starting from its sequence is central to structural and functional genomics, but it has been cast into particularly sharp focus when applied to signaling domains. For this review, we have chosen several examples of domain families whose structures and functions have recently been uncovered (Table 1). These cases illustrate broader trends in the synergy between the traditional hypothesis-driven paradigm of biochemical research and the more recent discovery-driven paradigm.

The examples in the table also illustrate how structural biology has had a dramatically increased presence in the early stages of understanding the function of newly discovered domains. The systematic structure determination of signaling domains as a class fits at least one of the definitions of structural genomics [3]. This brings us to the title of this article: 'Structural genomics and signaling domains'. The examples described in this article from our own group and from that of Shapiro were the result of explicitly taking a structural genomics approach (Fig. 1). The authors of the other studies do not describe their work explicitly in structural genomics terms, but we argue that approaches being taken, the targets chosen and the collective outcome of these studies are not so different from what might have been produced by a coordinated effort. As large-scale structural genomics initiatives begin in earnest, recent experiences with signaling domains offer hints about what might be in store as these initiatives move from early demonstration projects into areas with wide-ranging impacts on fundamental questions in biology and molecular medicine.

## Tubby

The tubby-like protein (TULP) family first came to light because, in mice, the mutation of the gene for

James H. Hurley\*  
D. Eric Anderson  
Bridgette Beach  
Bertram Canagarajah  
Yew Seng Jonathan Ho  
Eudora Jones  
Greg Miller  
Saurav Misra  
Matt Pearson  
Layla Saidi  
Silke Suer  
Ray Trievel  
Yosuke Tsujishita  
Laboratory of Molecular  
Biology, National Institute  
of Diabetes and Digestive  
and Kidney Diseases,  
National Institutes of  
Health, Bethesda,  
MD 20892-0580, USA.  
\*e-mail: jh8e@nih.gov



Table 1. Signaling domains with recently described structures<sup>a</sup>

| Domain    | Function <sup>b</sup>  | Biochemical function discovered from <sup>c</sup> | SMART sized <sup>d</sup> | Actual size <sup>e</sup> | Structure  |
|-----------|--|---|--------------------------|--------------------------|--|
| DEP       | Unknown  | NA  | 75                       | 94                       | 3-helix bundle plus $\beta$ -hairpin arm   |
| TULP core | PtdIns(4,5) $P_2$ -regulated transcription factor                          | Structural genomics                               | NA                       | 263                      | Novel fold with a $\beta$ -barrel filled by C-terminal $\alpha$ helix, basic groove for DNA binding  |
| START     | Lipid monomer binding and transport  | Structural genomics                               | 206                      | 229                      | Unclosed $\beta$ -barrel capped by a C-terminal helix, hollowed out core with a hydrophobic tunnel for lipid binding, similar fold to mammalian PTP and to plant allergens |
| ENTH      | PtdIns(4,5) $P_2$ and protein-protein interactions                         | Hypothesis-based                                  | 137                      | 149                      | Helical superhelix similar to VHS  |
| VHS       | Endocytic signal sequence binding  | Hypothesis-based                                  | 136                      | 153                      | Helical superhelix similar to ENTH   |
| PX        | Binds PtdIns(3) $P$ , other phosphoinositides                              | Hypothesis-based                                  | 118                      | 143                      | Novel $\alpha\beta$ fold   |
| PB1       | Protein-protein interaction with 'PC' motif in small G proteins and others | Hypothesis-based                                  | NA                       | 79                       | Similar to Ras-binding domain of Raf   |
| GAF       | Binds cGMP, chromophore, other small molecules                             | Inferred from larger proteins                     | 150                      | 180                      | Similar to PAS, another sensory and signaling domain   |
| IPP5C     | Phosphoinositide 5-phosphatase   | Inferred from larger proteins                     | 299                      | 336                      | Similar to DNaseI and DNA repair enzymes such as APE1  |

<sup>a</sup>Abbreviations: NA, not applicable; PtdIns(3) $P$ , phosphatidylinositol (3)-monophosphate; PtdIns(4,5) $P_2$ , phosphatidylinositol (4,5)-bisphosphate.  
<sup>b</sup>'Function' refers to the best-known function(s) for the domain group, and is not inclusive of all cases.  
<sup>c</sup>The approach indicated is that judged by the authors to be the most important single contributor to revealing the biochemical (as opposed to the cellular or genetic) function. In general, multiple approaches contributed. 'Inferred from larger proteins' is distinct from 'hypothesis-based' in that at least some of the functions of the former were established before the domain was identified as a conserved signaling motif.  
<sup>d</sup>The predicted size (number of amino acid residues) of the domain from SMART alignments, choosing a particular domain for which a structure has been determined.  
<sup>e</sup>The actual size (number of amino acid residues) of the domain as solved by X-ray crystallography or nuclear magnetic resonance, including essential extensions where present. These extensions do not appear to be part of the core fold of the domain in all cases, yet their inclusion is often essential to obtaining folded and functional protein. The differences between the predicted and actual sizes illustrate that substantial experimental effort is required to determine the correct boundaries of a predicted domain.

which the family is named, leads to maturity-onset obesity [4,5]. The TULP core domain comprises the conserved C-terminal portion of these proteins. Genetic and sequence information did not suggest a biochemical mechanism for the role of the TULP domain in obesity. Instead, structure determination proved to be the key to understanding the function of the TULP domain [6]. The TULP story is one of the most interesting contributions of structural genomics, combined with other discovery-driven assays, to enhancing our understanding of signaling mechanisms.

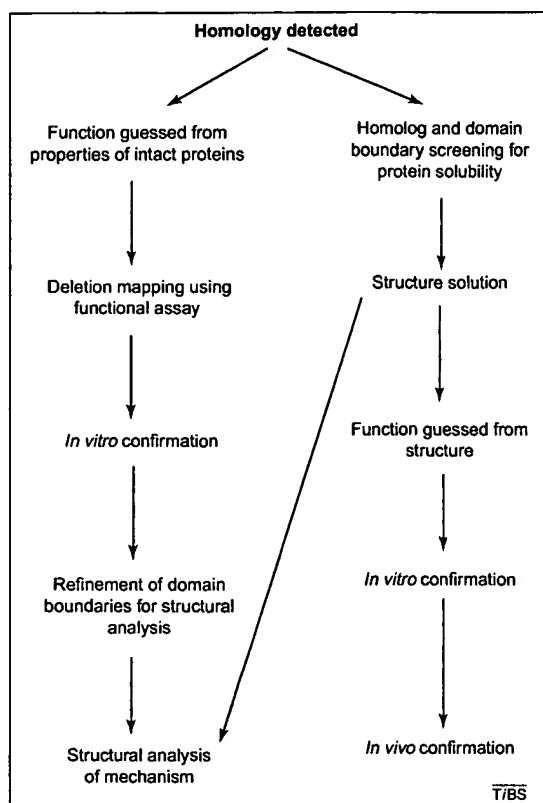
The tubby-TULP core domain has a novel fold comprising a 12-stranded  $\beta$ -barrel that completely encircles a hydrophobic  $\alpha$  helix that runs the length of the barrel [6]. The clue to the function of TULP came from a large curved basic surface with an appropriate size, shape and charge for binding DNA. Structure determination led to the hypothesis that the TULP core domain was the DNA-binding domain of a transcription factor, which was borne out by functional assays [6]. However, the question of the upstream regulation of tubby remained unanswered. This time, an answer was suggested by using the signal trap assay and by monitoring subcellular localization using green fluorescent protein (GFP) fusions. The tubby-TULP core domain localizes initially to the plasma membrane, but slowly dissociates and enters the nucleus [7]. This observation led to the hypothesis that lipid turnover at the plasma membrane drove membrane desorption, and again the hypothesis was confirmed by further experimentation, including structure determination of the TULP-PtdIns(4,5) $P_2$  headgroup complex.

### The DEP domain

The DEP (dishevelled-eglin-pleckstrin homology) domain is a widespread motif found in proteins involved in wnt signaling, regulators of G-protein signaling (RGS) proteins, pleckstrin and other signaling proteins [8]. Binding partners are not known for any DEP domain, but a role has been established for the Dishevelled DEP domain in wnt signaling based on mutations that interfere with biological function (reviewed in Ref. [9]). The DEP domains of several proteins have a membrane-targeting function [10,11], although the molecular mechanism of targeting is unknown.

If the tubby story represents the best-case scenario for structural genomics-based elucidation of function, the recent nuclear magnetic resonance (NMR) structural analysis of the Dishevelled DEP domain is perhaps a more typical illustration of what can and cannot be gleaned about a signaling domain from its three-dimensional structure alone [9]. The structure of this 100 amino acid domain consists of a three-helix bundle, a  $\beta$ -hairpin and two short C-terminal  $\beta$  strands. The analysis did not reveal any structural similarities to known structures that might provide clues as to the function of this domain. However, molecular surface features of Dishevelled-DEP did provide two clues as to the biochemical function of this domain. The first clue came from the observation of a cluster of seven basic residues that formed a flat patch on one side of the domain, which is typical of acidic phospholipid membrane-binding sites. This suggests that the membrane-targeting mechanism of DEP domains might involve direct binding of DEP domains to the negatively charged and roughly

**Fig. 1.** Two simplified paradigms for discovering the function of a newly identified domain, a function-based and hypothesis-driven approach (left), and a discovery-driven approach (right). The shunt from 'structure solution' to 'structural analysis of mechanism' illustrates just one aspect of the many levels of interplay between the different approaches, which are often being executed simultaneously by different laboratories.



planar surface of phospholipid bilayers in the cell. The second clue came from a mutation of a Lys residue that was already known to abrogate the function of Dishevelled in wnt signaling (reviewed in Ref. [9]). Lys434 is on a different face of the domain to the putative membrane-interacting face. This suggests that there are at least two functionally important interaction sites on the domain. The story of DEP domain function will remain incomplete, however, until binding partners of this domain are identified.

**START: lipid transporters, signaling proteins and more** START domains are found in a surprisingly diverse collection of proteins, including known and putative lipid transporters, transcription factors, enzymes of lipid metabolism, and signaling proteins [12]. The START domain is named after the steroidogenic acute regulatory (StAR) protein. The StAR protein is crucial for steroid hormone production because it is essential for the delivery of cholesterol to the inner membrane (IM) of mitochondria, where the first enzymatic reaction of steroidogenesis takes place. Mutations within the StAR-START domain cause congenital lipid adrenal hyperplasia (reviewed in Ref. [12]). Although the cellular role of StAR is well-known, the biochemical mechanism whereby cholesterol is delivered to the IM has been unclear. The recent crystal structure of the START domain of a cousin of StAR, MLN64, revealed a hollowed-out protein containing a hydrophobic tunnel big enough to bind one molecule of cholesterol and completely exclude it

from solvent [13] (Fig. 2). The structure-based hypothesis was confirmed by direct binding studies on the StAR and MLN64 START domains [13]. This suggests that the START domain functions in lipid transport by binding lipid monomers and sequestering them from solvent to deliver them across the aqueous compartments of the cell.

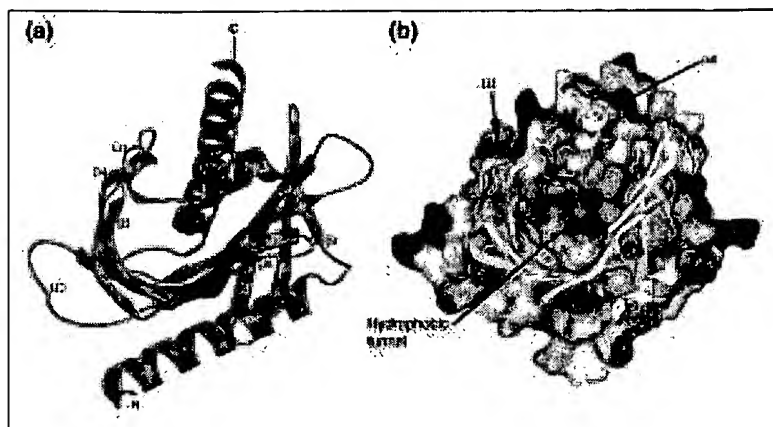
#### Taking it to the ENTH degree

The ENTH (Epsin N-terminal homology) [14] and VHS (Vps27, Hrs, STAM) [15] domains occur at the N-termini of proteins involved in intracellular trafficking. Crystal structures of Epsin-ENTH [16], Hrs-VHS [17] and Tom1-VHS [18] revealed that these two domains have very similar structures consisting of right-handed, eight-helical superhelices. The structures described in the first round of studies focused attention on, but fell short of answering, the question of the biochemical function of the ENTH and VHS domains. In the case of ENTH, the breakthrough came from structural and functional studies of the extended ENTH domains of two other endocytic proteins, AP180 [19] and CALM [20]. These proteins are regulated by PtdIns(4,5) $P_2$ , and studies of the N-terminal domains showed this regulation to be mediated by their ENTH domains. Crystal structures of the phosphoinositide complexes followed, revealing the binding sites, and also showing that these N-terminal domains are essentially an expanded version of the ENTH domain. Concurrently, the Epsin-ENTH was shown to weakly bind PtdIns(4,5) $P_2$  at a site near but not identical to the AP180 and CALM sites [21]. This discovery might not have been convincing taken in isolation, but makes sense in context with the AP180 and CALM results. The AP180 and CALM ENTH domains bind PtdIns(4,5) $P_2$  with moderate affinity at a site with several basic residues. The developing picture is that the ENTH domain family, similar to the pleckstrin homology (PH) domain family, is a collection of phosphoinositide-binding domains with widely varying affinity and specificity.

#### High fidelity: VHS domains

Despite the structural similarities of VHS and ENTH domains, they appear to serve distinct functions. Again, biological hypotheses drove the biochemical discoveries. A family of VHS domain-containing proteins called the GGAs (Golgi-localized  $\gamma$ -adaptin ear-domain-containing) was recently reported to direct intracellular trafficking in the endosomal-lysosomal pathway. With the other domains accounting for functions in Arf and clathrin binding, the VHS was the only remaining orphan domain in the GGAs. Known orphan cargo proteins then became candidates for interactions with the GGA VHS domains. This approach revealed that the function of the GGA-VHS domain is to recognize an acidic-cluster-dileucine signal within the cytoplasmic C-terminal tails of membrane proteins that are trafficked from the *trans* Golgi network to





**Fig. 2.** (a) The START domain fold, an unclosed  $\beta$ -barrel capped a C-terminal helix. (b) Sagittal slice through the molecular surface of the MLN64 START domain shown in the same perspective as in (a). The surface cutaway reveals a mostly hydrophobic tunnel running through the center of the protein. The observation of a hydrophobic tunnel in the middle of this domain was the central observation leading to the proposal that START domain proteins bind and transport lipid monomers. The surface is colored blue (basic), red (acidic), green (hydrophobic) and white (uncharged polar). The interior of the protein is gray. Figure is adapted, with permission, from Ref. [13].

endosomes [22–25]. The structures of GGA-VHS domains bound to signal peptides were solved in quick succession, and revealed the signal peptide binding sites (S. Misra *et al.*, unpublished).

#### PX

PX domains attracted attention in two waves. The first wave hit when their presence was noted in a large number of signaling proteins, especially the NADPH oxidase phox subunits for which the domain is named [26]. This was followed, after several years of latency, by a small tsunami of attention when four groups simultaneously reported that certain PX domains are cellular receptors for phosphatidylinositol 3-phosphate [PtdIns(3)P] [27–31]. Various other PX domains appear to bind other phosphoinositides [30,32]. The intellectual basis for the discovery resembled the case for ENTH: several PX-domain-containing proteins – Vam7p, SNX3, and the p40 and p47 subunits of the NADPH oxidase complex – were known to translocate in response to PI 3-kinase signaling. The PX domain became a natural candidate for the PtdIns(3)P-responsive module within these proteins, and from there it was a matter of pinning down the details. The NMR structure of the PX domain of p47<sup>phox</sup> was determined before the phosphoinositide-binding function of the PX domain had been established [33]. This study focused attention on the interaction of the PX domain with the SH3 domain of the same protein. Such intraprotein, interdomain interactions are undoubtedly important for allosteric regulation, although that is not usually what is construed when ‘the function’ of a domain is discussed. Fortunately, the second structure of a PX domain, following on the heels of the discovery of the function, was an X-ray study of the complex of the p40<sup>phox</sup>-PX with a short-chain PtdIns(3)P [34].

#### PB1

The PB1 motif is a newly described domain involved in interacting with the small G protein Cdc42p [35]. The trend towards the compression of events in bioinformatics, structure and function, was highlighted by two back-to-back reports on the discovery, function and structure of the PB1 domain [35,36]. The function of the Bem1p-PB1 domain was established in a straightforward way, by deletion mapping of the determinants for the already-known interaction with Cdc42p. The NMR structure of this domain [36] revealed similarity to the Ras-association domain of Raf, consistent with its function in the binding of small G proteins.

#### GAF

GAF domains [37] are among the most widespread of all signaling domains, found in 826 proteins in the SMART database. Cyclic GMP (cGMP)-regulated cyclic nucleotide phosphodiesterases (PDEs) contain GAF domains, which are the allosteric binding sites for cGMP. Hundreds of other signaling proteins also contain GAF domains, including essential plant signaling enzymes such as the photosensing phytochromes and the ethylene receptor, and a vast array of microbial signaling and sensory proteins. Thus, the functions of a handful of GAF domains are known, including the cGMP-binding GAF domains of the PDEs, and the chromophore-binding GAF domains of the phytochromes. Remarkably, the functions of most GAF domains remain unknown, and no structure of a GAF domain was known until last year.

The crystal structure of a GAF domain from a yeast protein of unknown function, YKG9, revealed a close structural similarity to another very widespread class of signaling and sensory domain, the PAS (Per Arnt Sim) domain [38]. Three PAS domain structures have been solved, and two out of the three bind a chromophore or a heme in a distinctive buried pocket on one side of the central  $\beta$  sheet. The GAF domain structure revealed an unusual buried pocket that coincides structurally with the heme- or chromophore-binding pocket of the PAS domain. Thus, the GAF and PAS domains together form a structurally, and almost certainly evolutionarily, related family of small-molecule-binding domains. What the structure does not reveal is the nature of the small molecule that binds in the pocket, either of the YKG9 GAF domain, or of other GAF domains. On occasion, structural genomics has succeeded in identifying small molecules and cofactors that bind to proteins of previously unknown function, as these ligands can be present in the host cells used for recombinant protein expression. This sort of strategic serendipity is most likely to occur when the expressing host cell is similar to the organism in which the protein occurs naturally, and cannot be counted upon when eukaryotic protein domains are expressed in a prokaryotic host.

The most pressing question for understanding the biological function of the large majority of GAF domain proteins is to determine the identities of the small-molecule ligands of these GAF domains. This remains a challenging objective, as no general method for doing this exists.

#### Inositol polyphosphate 5-phosphatases

The inositol polyphosphate 5-phosphatases have been intensively studied over the past decade because of their profound importance in an enormous range of cellular processes [39]. These enzymes, and their catalytic (IPP5C) domains, might seem to be in odd company among the domains described above, most of which were only very recently discovered and studied. Despite efforts by many groups over the years, no crystal structure of any well-characterized IPP5C domain has been obtained. The use of a structural genomics-inspired tactic, database searching for previously unstudied homologs that might be susceptible to crystallization, led to the cloning, characterization, and crystal structure determination of a new member of this family, a previously unnamed and uncharacterized protein from *S. pombe* that is now known as SPsynaptojanin [40]. Although the function of this domain was well-established, the structure led to new insights into the catalytic mechanism and substrate specificity, an example of the interplay between structural genomic and traditional structural biology approaches.

#### Conclusions

The studies described in this article illustrate some of the insights that have been obtained into a range of biological processes by approaches that focus on protein domains. It is worth considering the relative contributions of structural and functional approaches, and of hypothesis-driven versus discovery-driven paradigms. Different approaches represent a continuum between these poles, and it is the interplay between the approaches that is most revealing.

The tubby study is a vivid illustration of the power of a discovery-driven and structural genomic approach to function. The discovery that tubby is a transcription factor was driven by the observation of a large basic region on the structure that literally 'looked like' a DNA-binding site. Subsequent hypothesis-driven functional assays were essential to confirm this idea. The second tubby breakthrough was propelled by a different type of unbiased discovery-driven assay using GFP fusions. Similar to tubby, involvement of the StAR protein was implicated by genetics both in disease and in regulation of normal physiology, yet its biochemical function was unclear. The clues to function were more extensive than for tubby, but again, it was the crystal structure that was pivotal in terms of suggesting mechanism.

The ENTH, VHS and PX domain stories illustrate the undiminished power of a good, old-fashioned hypothesis. In the ENTH and PX cases, full-length proteins containing these domains were known to be involved in PtdIns(4,5) $P_2$ - and PtdIns(3) $P$ -dependent processes, respectively. The discovery of the function of the VHS domain had to wait for the discovery of a new class of VHS-domain-containing protein, the GGAs.

The first structures of ENTH, VHS, PX and DEP domains were determined before the elucidation of their biological functions, and were not particularly enlightening in a functional sense. What they did provide, at least for ENTH, VHS and PX (the function of the DEP domain remained unknown) was a powerful impetus for a second round of much more informative structural studies carried out in the full light of the known function. The second wave followed so closely on the heels of the first reports of known function partly because of the groundwork laid by the initial wave. Although structure might not always reveal function by itself, in the absence of pre-existing biochemical work, the structure of a novel domain casts what was heretofore a purely bioinformatic construction, into physical and chemical 'reality' for the first time.

What makes some structures more revealing about function than others? Bacterial structural genomics has had some notable success in predicting function from structure. Many of these successes have involved enzymes, in which fold similarity and conserved catalytic geometry provide powerful insights. Fold similarity is a powerful predictor of common function in these cases, but where ligand-binding domains are concerned, it might be much less useful. For example, the structural similarity between the ENTH and VHS domains did not prove to be a major factor in our understanding of their function. In many cases, bacterial enzymes expressed in a bacterial host are purified in complex with relevant cofactors. When a eukaryotic protein, to take YKG9-GAF as an example, is expressed in a prokaryotic host, there is no guarantee that the relevant cofactor or modification will be present.

In eukaryotes, where regulatory and signaling proteins outnumber metabolic proteins, identification of protein, nucleic acid, membrane lipid and small-molecule ligands is probably important in more cases than is identifying a novel enzyme activity. The primary function of many eukaryotic signaling domains is their binding to other proteins. For these cases, other scaleable approaches such as the yeast two-hybrid and pulldown or proteomics methods will be more suitable. It is often possible to spot a nucleic acid, phospholipid membrane, or hydrophobic small-molecule ligand-binding site by inspection of a structure, and these situations favor structural genomics. In summary, the question posed at the

outset of this article: 'what is the function of the FITB domain?', is difficult enough to answer to warrant the implementation of all available approaches.

Domains will probably represent the bulk of protein targets in eukaryotic structural genomics because so many intact eukaryotic proteins are large enough to be challenging to express and crystallize. The crucial difference between studying novel domains and intact proteins is that the correct boundaries need to be established for each new domain. In the past, the function of a protein domain was usually established before the structural work began. Functional analysis usually involves deletion mapping that produces at least a rough starting point for expression of protein for structural studies. Without the benefit of a functional assay, physical-chemical properties of the domain (typically meaning its solubility) are the only basis for the assay.

This significantly increases the effort involved, but not to a prohibitive degree. It is encouraging that so many domain structures (e.g. the first structures for ENTH, VHS, PX and DEP) are being produced where the solubility of the recombinant protein is the sole guide to the choice of boundaries.

If genome sequences are the 'parts list of life', one could argue that the sum total of signaling domains, together with their binding partners and post-translational modifications, represent a 'parts list for signal transduction'. It is gratifying to see the convergence of bioinformatic, structural and functional techniques, and the interplay between hypothesis- and discovery-driven paradigms. As attention in signal transduction shifts towards network models, the accumulation of domain data should contribute to quantitative and predictive models for cell signaling.

## References

- Bork, P. *et al.* (1997) Cytoplasmic signalling domains: the next generation. *Trends Biochem. Sci.* 22, 296–298
- Schultz, J. *et al.* (2000) SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* 28, 231–234
- Terwilliger, T.C. *et al.* (1998) Class-directed structure determination: foundation for a protein structure initiative. *Protein Sci.* 7, 1851–1856
- Kleyn, P.W. *et al.* (1996) Identification and characterization of the mouse obesity gene tubby: a member of a novel gene family. *Cell* 85, 281–290
- NobenTrauth, K. *et al.* (1996) A candidate gene for the mouse mutation tubby. *Nature* 380, 534–538
- Boggon, T.J. *et al.* (1999) Implication of tubby proteins as transcription factors by structure-based functional analysis. *Science* 286, 2119–2125
- Santagata, S. *et al.* (2001) G-protein signaling through tubby proteins. *Science* 292, 2041–2050
- Ponting, C.P. and Bork, P. (1996) Pleckstrin's repeat performance: a novel domain in G-protein signaling? *Trends Biochem. Sci.* 21, 245–246
- Wong, H.C. *et al.* (2000) Structural basis of the recognition of the Dishevelled DEP domain in the Wnt signalling pathway. *Nat. Struct. Biol.* 7, 1178–1184
- Axelrod, J.D. *et al.* (1998) Differential recruitment of Dishevelled provides signaling specificity in the planar cell polarity and Wingless signaling pathways. *Genes Dev.* 12, 2610–2622
- Koelle, M.R. and Horvitz, H.R. (1996) EGL-10 regulates G protein signaling in the *C. elegans* nervous system and shares a conserved domain with many mammalian proteins. *Cell* 84, 115–125
- Ponting, C.P. and Aravind, L. (1999) START: a lipid binding domain in StAR, HD-ZIP and signalling proteins. *Trends Biochem. Sci.* 24, 130–132
- Tsujishita, Y. and Hurley, J.H. (2000) Structure and lipid transport mechanism of a StAR-related domain. *Nat. Struct. Biol.* 7, 408–414
- Kay, B.K. *et al.* (1999) Identification of a novel domain shared by putative components of the endocytic and cytoskeletal machinery. *Protein Sci.* 8, 435–438
- Lohi, O. and Lehto, V.P. (1998) VHS domain marks a group of proteins involved in endocytosis and vesicular trafficking. *FEBS Lett.* 440, 255–257
- Hyman, J. *et al.* (2000) Epsin 1 undergoes nucleocytoplasmic shuttling and its Eps15 interactor NH2-terminal homology (ENTH) domain, structurally similar to armadillo and HEAT repeats, interacts with the transcription factor promyelocytic leukemia Zn2+ finger protein (PLZF). *J. Cell Biol.* 149, 537–546
- Mao, Y.X. *et al.* (2000) Crystal structure of the VHS and FYVE tandem domains of Hrs, a protein involved in membrane trafficking and signal transduction. *Cell* 100, 447–456
- Misra, S. *et al.* (2000) Structure of the VHS domain of human Tom1 (target of myb 1): insights into interactions with proteins and membranes. *Biochemistry* 39, 11282–11290
- Mao, Y.X. *et al.* (2001) A novel all helix fold of the AP180 amino-terminal domain for phosphoinositide binding and clathrin assembly in synaptic vesicle endocytosis. *Cell* 104, 433–440
- Ford, M.G.J. *et al.* (2001) Simultaneous binding of PtdIns(4,5)P<sub>2</sub> and clathrin by AP180 in the nucleation of clathrin lattices on membranes. *Science* 291, 1051–1055
- Itoh, T. *et al.* (2001) Role of the ENTH domain in phosphatidylinositol-4,5-bisphosphate binding and endocytosis. *Science* 291, 1047–1051
- Nielsen, M.S. *et al.* (2001) The sortilin cytoplasmic tail conveys Golgi-endosome transport and binds the VHS domain of the GGA2 sorting protein. *EMBO J.* 20, 2180–2190
- Puertollano, R. *et al.* (2001) Sorting of mannose 6-phosphate receptors mediated by the GGAs. *Science* 292, 1712–1716
- Zhu, Y.X. *et al.* (2001) Binding of GGA2 to the lysosomal enzyme sorting motif of the mannose 6-phosphate receptor. *Science* 292, 1716–1718
- Takatsu, H. *et al.* (2001) GGA proteins interact with acidic dileucine sequences within the cytoplasmic domains of sorting receptors through their VHS domains. *J. Biol. Chem.* 276, 28541–28545
- Ponting, C.P. (1996) Novel domains in NADPH oxidase subunits, sorting nexins, and PtdIns 3-kinases: binding partners of SH3 domains? *Protein Sci.* 5, 2353–2357
- Wishart, M.J. *et al.* (2001) Phox lipids: revealing PX domains as phosphoinositide binding modules. *Cell* 105, 817–820
- Cheever, M.L. *et al.* (2001) Phox domain interaction with PtdIns(3)P targets the Vam7 t-SNARE to vacuole membranes. *Nat. Cell Biol.* 3, 613–618
- Xu, Y. *et al.* (2001) SNX3 regulates endosomal function through its PX-domain-mediated interaction with PtdIns(3)P. *Nat. Cell Biol.* 3, 658–666
- Kanai, F. *et al.* (2001) The PX domains of p47phox and p40phox bind to lipid products of PI(3)K. *Nat. Cell Biol.* 3, 675–678
- Ellson, C.D. *et al.* (2001) PtdIns(3)P regulates the neutrophil oxidase complex by binding to the PX domain of p40phox. *Nat. Cell Biol.* 3, 679–682
- Song, X. *et al.* (2001) Phox homology domains specifically bind phosphatidylinositol phosphates. *Biochemistry* 40, 8940–8944
- Hiroaki, H. *et al.* (2001) Solution structure of the PX domain, a target of the SH3 domain. *Nat. Struct. Biol.* 8, 526–530
- Bravo, J. *et al.* (2001) The crystal structure of the PX domain from p40phox bound to phosphatidylinositol 3-phosphate. *Mol. Cell* 8, 829–839
- Ito, T. *et al.* (2001) Novel modular domain PB1 recognizes PC motif to mediate functional protein-protein interactions. *EMBO J.* 20, 3938–3946
- Terasawa, H. *et al.* (2001) Structure and ligand recognition of the PB1 domain: a novel protein module binding to the PC motif. *EMBO J.* 20, 3947–3956
- Aravind, L. and Ponting, C.P. (1997) The GAF domain: an evolutionary link between diverse phototransducing proteins. *Trends Biochem. Sci.* 22, 458–459
- Ho, Y.S.J. *et al.* (2000) Structure of the GAF domain, a ubiquitous signaling motif and a new class of cyclic GMP receptor. *EMBO J.* 19, 5288–5299
- Majerus, P.W. *et al.* (1999) The role of phosphatases in inositol signaling reactions. *J. Biol. Chem.* 274, 10669–10672
- Tsujishita, Y. *et al.* (2001) Specificity determinants in phosphoinositide dephosphorylation: crystal structure of an archetypal inositol polyphosphate 5-phosphatase. *Cell* 105, 379–389

# CDD: a database of conserved domain alignments with links to domain three-dimensional structure

Aron Marchler-Bauer\*, Anna R. Panchenko, Benjamin A. Shoemaker, Paul A. Thiessen, Lewis Y. Geer and Stephen H. Bryant

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38 A, Room 8N805, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received August 15, 2001; Accepted October 1, 2001

## ABSTRACT

The Conserved Domain Database (CDD) is a compilation of multiple sequence alignments representing protein domains conserved in molecular evolution. It has been populated with alignment data from the public collections Pfam and SMART, as well as with contributions from colleagues at NCBI. The current version of CDD (v.1.54) contains 3693 such models. CDD alignments are linked to protein sequence and structure data in Entrez. The molecular structure viewer Cn3D serves as a tool to interactively visualize alignments and three-dimensional structure, and to link three-dimensional residue coordinates to descriptions of evolutionary conservation. CDD can be accessed on the World Wide Web at <http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml>. Protein query sequences may be compared against databases of position-specific score matrices derived from alignments in CDD, using a service named CD-Search, which can be found at <http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>. CD-Search runs reverse-position-specific BLAST (RPS-BLAST), a variant of the widely used PSI-BLAST algorithm. CD-Search is run by default for protein–protein queries submitted to NCBI's BLAST service at <http://www.ncbi.nlm.nih.gov/BLAST>.

## INTRODUCTION

Protein domains may be thought of as proteins' structural and functional building blocks, dividing the primary and tertiary structure of a chain into distinct units. Domains are also mobile genetic units, rearranging in various combinations throughout the molecular evolution of proteins. To understand such processes, and the effect they have had on the present protein repertoire, proteins need to be analyzed not as full-length sequences but rather as collections of individual domains, each of which is important as a unit of molecular evolution.

Protein sequence comparison, as a tool to investigate patterns of conservation and divergence in molecular evolution, is more powerful when sequences are compared to models of protein families instead of other single sequences (1). This may be of particular importance if one wants to use sequence comparison

for domain identification and annotation of new sequence data. Examples of such family models are protein profiles (2), hidden Markov models (3,4) and position-specific score matrices (5). The latter, for example, are constructed automatically from sets of pairwise alignments in an iterative database search procedure in PSI-BLAST, a popular addition to the BLAST family of programs (6). PSI-BLAST generates position-specific score matrices, anchored on query sequences, to serve as models of protein families. Explicit alignment models, though, are important for further analysis of divergent domain families and for the transfer of annotation.

Collections of domain alignment models thus are invaluable resources for the study of protein evolution and for large-scale annotation of genomic sequences. Examples of carefully compiled collections are Pfam (7) and SMART (8), which also come bundled with powerful, hidden Markov model-based search engines. Alignments from Pfam and SMART have been imported into the Conserved Domain Database (CDD), to serve a variety of purposes. CDD lets us tie explicit alignment models to a fast search system using BLAST heuristics, via position-specific score matrices computed from the alignments. A corresponding search service has been made available at <http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>. Also, CDD makes possible the interactive visualization of both multiple alignment data and protein three-dimensional structure, using NCBI's Cn3D viewer (9). Such combined alignment/structure displays are available for both the browsing of CDD content, and for the display of database search results.

## THE CONSERVED DOMAIN DATABASE

We have set up a protocol to import and present alignment data from external sources as well as from in-house collaborators. We attempt to identify the sequence fragments used by the alignments' authors, so that we can link to full-length sequences in Entrez (10). If accession codes supplied by the source databases cannot be identified, BLAST searches are run for the fragments in order to find identical or very similar sequences in NCBI's databases, requiring at least 90% sequence identity across the aligned fragment. Particular attention is paid to close matches with structure-linked sequences, and we substitute alignment rows with such sequences when possible. For substitution, we require a similarity threshold of at least 75% sequence identity in the aligned region and no more

\*To whom correspondence should be addressed. Tel: +1 301 435 4919; Fax: +1 301 480 9241; Email: [bauer@ncbi.nlm.nih.gov](mailto:bauer@ncbi.nlm.nih.gov)

than 5% of that region is allowed to be lost due to insertions and deletions.

Upon import, multiple alignments are deconstructed into sets of pairwise alignments. A representative common to all the pairwise alignments is chosen as the sequence with the fewest deletions relative to other sequences, so that the loss of alignment information is minimal. In fact most of the alignments imported from Pfam or SMART have a very pronounced block structure, reducing the risk of losing information in this step. Structure-linked sequences are picked as representatives whenever available.

Imported domain alignments can be retrieved by accession or searched by keyword at <http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml>. The server generates summary pages from which several alignment visualization styles are available. If three-dimensional structure information is available, Cn3D 3.0, a molecular structure viewer distributed by NCBI, can be used to display integrated views of the domain's multiple alignment and its conservation patterns, as well as the three-dimensional structure of a representative member. This display allows interactive highlighting and feature annotation. Figure 1 shows an example of how this capability can be used to illustrate how genotypes may be linked to disease.

Domain alignments in CDD are used to calculate position-specific score matrices for database searching. For the representation of position-specific score matrix (PSSM) models, a consensus sequence is calculated for each conserved domain, reporting the most frequent residues in aligned columns. Although visible in alignment displays, the consensus sequence is not used directly in PSSM calculations. However, it determines the length of the PSSMs, as only columns with >50% aligned states are included in the consensus and PSSM calculation.

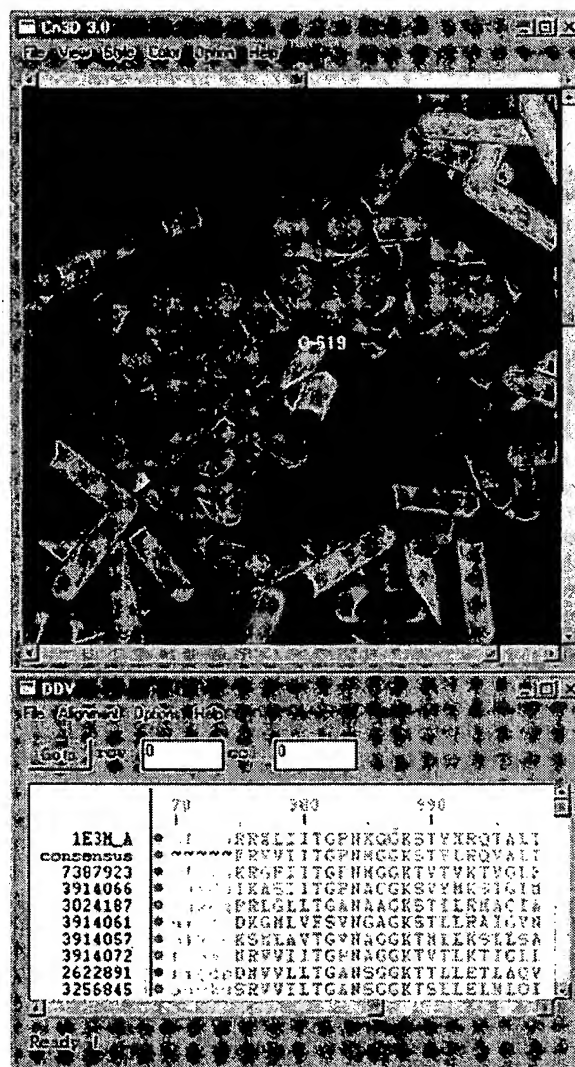
The search engine making use of CDD's collection of PSSM models is reverse-position-specific BLAST (RPS-BLAST), a variant of PSI-BLAST. It inverts the role of query and subject, comparing a single sequence against a database of PSSM models instead of searching a database of sequences with a single PSSM model. A web-based interface to RPS-BLAST, CD-Search, is available at <http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>.

The ASN.1 data specification for domain multiple alignment data in CDD is available through the NCBI toolbox distribution at <ftp://ncbi.nlm.nih.gov/toolbox>, together with C program code that can be used to read, write and compute with CDD data in the context of the NCBI toolkit. The content of the CDD can be downloaded from NCBI's FTP site in machine-readable ASN.1 format, by following instructions on the CDD home page.

## FUTURE DEVELOPMENTS

DART, the domain architecture retrieval tool, is an application making use of CDD data. It runs a variant of CD-search for protein query sequences and compares the inferred domain architecture of the query with pre-calculated domain architectures of database proteins, showing a list of neighbors with matching sets of domains. DART can be accessed at <http://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi?cmd=rps>.

The DART service covers a non-redundant subset of Entrez's protein database only. In an ongoing effort, CDD will



**Figure 1.** Cn3D 3.0 view showing a subset of aligned sequences from the CD sm (ATPase domain of DNA mismatch repair MUTS family). Residues corresponding to the P-loop motif around the ADP/Mg<sup>2+</sup> binding site have been annotated in Cn3D and are highlighted in green: GLY614-x(4)-GLY619-LYS620-SER621 in 1E3M chain A (12). The ADP/Mg<sup>2+</sup> complex is colored magenta. In the related human MSH2 protein, a somatic GLY→SER mutation in the position corresponding to GLY619 (highlighted in yellow) has been associated with type I familial non-polyposis colon cancer. From analyzing the image one may understand how a substitution of the side chain at this position interferes with ADP/Mg<sup>2+</sup> binding and may therefore impede the DNA mismatch repair system.

be made a fully integrated part of NCBI's Entrez system, making use of Entrez's powerful search and query refinement engine. As a prerequisite, all proteins in Entrez will be neighbored to CDD, thus adding domain annotations to all protein sequences, annotations which will be refreshed periodically, as both Entrez-protein and CDD will continue to grow. Conserved domain models in Entrez will not only be linked to proteins, but also to three-dimensional structure data, literature, nodes in the taxonomic classification and to other conserved domain

models, using suitable definitions for conserved domain neighbor relationships.

We have started to curate multiple alignments in CDD. Our goal is the reconciliation of sequence alignment data with quantitative information from protein three-dimensional structure and structure comparison, resulting in sets of pairwise alignments between each family member and a structure-linked representative that could be used to instantiate initial three-dimensional models for family members. One of the prerequisites is, for example, that structures inferred from alignments do not violate basic geometric principles. We also plan to validate CDD alignments with sequence-structure threading methods (11) as a means to detect errors in the alignments, outliers requiring manual curation and contamination with sequences from outside the family. Setting up a curation pipeline will allow us to periodically update conserved domain alignments with new family members, as sequence databases continue to grow. Carefully curated alignments will also serve as a means to link functional annotation to conserved residues, and to make this annotation accessible in visualization services. A set of curated conserved domains will be available later this year.

## ACKNOWLEDGEMENTS

We are grateful to the authors of Pfam and SMART, for creating an invaluable resource. We thank Chris Ponting, Alex Bateman, Eugene Koonin and David Lipman for discussions and helpful suggestions, L. Aravind for providing sequence alignment data, Tom Madden and Sergei Shavirin for making RPS-BLAST available, Richard Copley for providing access to SMART data, and Naomi Ariel for help with the initial analysis of imported alignments.

## REFERENCES

1. Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. and Chothia, C. (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.*, **284**, 1201–1210.
2. Gribskov, M., McLachlan, A.D. and Eisenberg, D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.
3. Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
4. Karplus, K., Barrett, C. and Hughey, R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.
5. Henikoff, S. and Henikoff, J.G. (1997) Embedding strategies for effective use of information from multiple sequence alignments. *Protein Sci.*, **6**, 698–705.
6. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
7. Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L. and Sonnhammer, E.L.L. (2000) The Pfam proteins family database. *Nucleic Acids Res.*, **28**, 263–266. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 276–280.
8. Schultz, J., Copley, R.R., Doerks, T., Ponting, C.P. and Bork, P. (2000) SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.*, **28**, 231–234. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 242–244.
9. Wang, Y., Geer, L.Y., Chappey, C., Kans, J.A. and Bryant, S.H. (2000) Cn3D: sequence and structure views from Entrez. *Trends Biochem. Sci.*, **25**, 300–302.
10. Wheeler, D.L., Church, D.M., Lash, A.E., Leipe, D.D., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Tatusova, T.A., Wagner, L. and Rapp, B.A. (2001) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **29**, 11–16. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 13–16.
11. Panchenko, A.R., Marchler-Bauer, A. and Bryant, S.H. (1999) Threading with explicit models for evolutionary conservation of structure and sequence. *Proteins*, **37** (Suppl. 3), 133–140.
12. Lamers, M.H., Perrakis, A., Enzlin, J.H., Winterwerp, H.H., de Wind, N. and Sixma, T.K. (2000) The crystal structure of DNA mismatch repair protein MutS binding to a G × T mismatch. *Nature*, **407**, 711–717.

**This Page is Inserted by IFW Indexing and Scanning  
Operations and is not part of the Official Record**

**BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ **BLACK BORDERS**
- ☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☐ **FADED TEXT OR DRAWING**
- ☐ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☒ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☒ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER:** \_\_\_\_\_

**IMAGES ARE BEST AVAILABLE COPY.**

**As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.**